Theorems that are not probability

Dr Nic Freeman

July 13, 2023

Contents

Chapter 0

Introduction

These notes cover various theorems in mathematics that are not about probability, but which can be proved using probabilistic techniques. Sometimes this provides an easier proof than the standard way, sometimes not, and sometimes probability *is* the standard way.

The range of theorems that fall into this category is rather impressive. Several of the most important parts of mathematics are mentioned on the contents page. I hope you'll recognize most of them. This provokes the question: *why* is probability apparently able to stick its nose into so many unrelated topics? I think there are two major effects at works here.

- Firstly, probability is the part of mathematics that 'owns' the concept of independence. This concept is incredibly useful: we find it intuitive to work with *and* it allows us to break down complicated systems into smaller, more manageable parts, whilst preserving an understanding of how those parts interact. Most of the probabilistic techniques that we'll use in the following chapters involve independence assumptions.
- Secondly, probability often acts as an extra layer that one can apply to deterministic objects, resulting (of course) in random objects. The objects themselves can be pretty much anything, and this allows probability interact with a wide range of mathematical areas. The usual boundaries that create subject areas in mathematics, such as continuous vs discrete or applied vs pure, do not get in the way here.

Probability was somewhat disregarded by mathematicians up until roughly the last 200 years. This was in part for idealistic reasons – it was associated with gambling, which was viewed as an unsavoury past-time. A more practical difficulty was that most of probability theory relies heavily on real analysis, and the rigorous basis for real analysis is also less than 200 years old. The modern foundations of probability theory date from the 1930s, which is very young for a large field of mathematics. This meant that probability both developed and grew rapidly during the 20th century, with the result that many important results in mathematics were already known via other means well before anyone thought about trying to prove them using probability.

I do not wish to suggest that probability should be any more prominent within mathematics than it already is. We have already reached a point at which most areas of mathematics use probabilistic techniques, some to a great extent and others rarely. The vast majority of probabilists are interested in probability – in constructing random objects and associated theories that might become useful to modellers – and are not trying to re-prove what has already been achieved elsewhere. Nonetheless, it is interesting to catalogue examples of when probability connects to something unexpected, because *any* unexpected connection might lead to a new idea, and new ideas are valuable.

It is also interesting from the point of view of learning mathematics. We tend to teach mathematics courses in (relative) isolation from each other[1], which is both efficient and practical. When proofs come from outside of familiar ground they can easily feel counter-intuitive and adventurous. They can also feel incredibly sneaky and clever, like having a map that takes an invisible shortcut; at least, for part of the way. I suspect that these effects are mostly artificial, as a side effect of how we learn, but that doesn't stop them from being fun.

On practical considerations: these notes are aimed at undergraduates in mathematics who are between their penultimate and final year. This provides a large amount of theoretical background to work from, but it also means taking a small number of (hopefully natural) things on trust within the last two chapters.

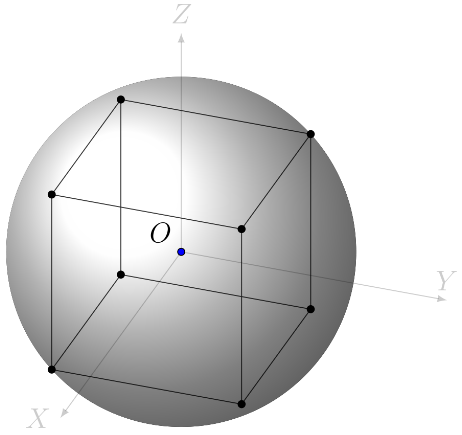
There are five exercises, one in each of Chapters 1-5. I hope these will be rather difficult and I have no plans to provide solutions. In some cases you will be able to find hints/solutions via searching online or within the textbooks, and I see nothing wrong with doing that if (and when) you get stuck.

[1] In fact, mathematics courses are much more tightly interwoven with each other (i.e. with pre-requisites) than in most other disciplines. Those of us who design mathematics degrees spend a non-trivial fraction of our time trying to explain this phenomenon to the rest of the university. By 'isolation' I mean only that concurrent courses tend not to depend upon on each other, despite having a large dependency on previous years.

Chapter 1

Existence results

Recall the meaning of *inscribing* a cube within a sphere: the cube sits inside of the sphere, with each vertex of the cube positioned on the surface of the sphere. It is not especially easy to illustrate this on a two dimensional page, but we can try:



Here's a question that, at first sight, is not about probability.

Problem 1.1 *Take a sphere of radius* 1*, and a cube of radius* 1*. Ten percent of the surface of the sphere is coloured red and the rest is coloured black. Is it possible to inscribe the cube inside the sphere, such that all vertices of the cube are coloured black?*

Perhaps it depends on which bits are coloured red?

There are two standard ways to construct things in mathematics. The first is to construct something explicitly, which is often difficult, particularly if the object in question is complicated, or you don't know exactly what you want. I don't know of an explicit solution to Problem 1.1. The second way is to use an implicit construction, which is to say that you're going to make some sort of equation, show that it has a solution, and that this solution will be the thing you want. This is very effective but, unfortunately, I can't solve Problem 1.1 that way either.

The 'probabilistic method' for proving existence sits somewhere in the middle. The idea is that you make a random object that *with some probability* will be the object you want. It doesn't matter how small that probability is; as long as the probability is strictly greater than zero then the object must exist. This allows you to try explicit constructions that have some chance of going wrong, and *this* works well for Problem 1.1.

Solution of Problem 1.1: Take a cube inscribed within the sphere, orientated uniformly at random. Strictly speaking, in three-dimensional spherical coordinates (r, θ, ϕ) this means we let θ and ϕ be independent uniform random variables on $(0, 2\pi)$. More importantly, it means that the location of a given vertex of the cube is distributed uniformly on the surface of the sphere.

Let X be the number of corners that are black. Label the corners from $i = 1, \ldots, 8$. We can write

$$X = \sum_{i=1}^{8} \mathbb{1}_{\{A_i \text{ is black}\}},$$

where A_i is the colour of the i^{th} corner. Here $\mathbb{1}_{\{A_i \text{ is black}\}}$ is equal to 1 if A_i is black and equal to zero if A_i is red. Hence,

$$\mathbb{E}[X] = \sum_{i=1}^{8} \mathbb{E}\left[\mathbb{1}_{\{A_i \text{ is black}\}}\right] = \sum_{i=1}^{8} \mathbb{P}[A_i \text{ is black}].$$

Since A_i is uniformly distributed on the surface of the sphere, and 90% of the sphere is black, we have $\mathbb{P}[A_i \text{ is black}] = \frac{9}{10}$. Hence,

$$\mathbb{E}[X] = \tfrac{9}{10} \times 8 = 7.2.$$

Note that X can only take the values $\{1, 2, \ldots, 8\}$. Since $\mathbb{E}[X] > 7$ and X takes integer values, we must have $\mathbb{P}[X = 8] > 0$. Therefore, there are orientations of the cube for which all 8 vertices are black. ■

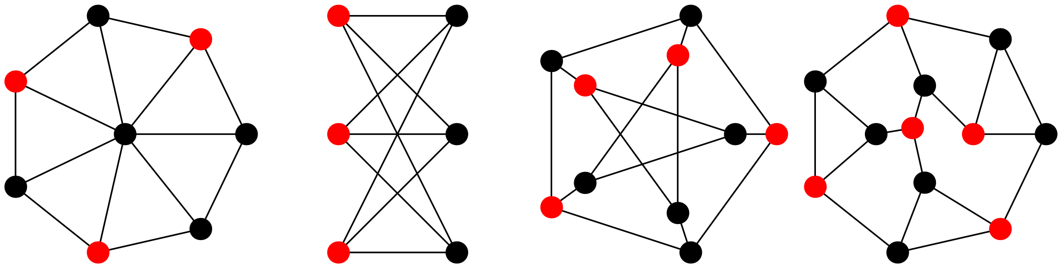
The solution is cute but it doesn't give us much of a sense of what the probabilistic method of construction can do. In retrospect it's not that impressive, because the proof makes clear that guessing at random would have solved the problem with fairly high probability. If the probability of finding a good configuration had been small then the trick with $\mathbb{E}[X]$ in the last paragraph wouldn't have worked.

In fact, probabilistic construction is a well established tool in discrete mathematics, which is used to prove existence of (mostly) finite graphs and related structures with particular properties. You will come across it in most third/fourth year graph theory or combinatorics courses. There are some clever ways to sharpen the technique and better cope with low probability events, and in the rest of this chapter we'll look at two examples of this.

1.1 Stable sets

Let $G = (V, E)$ be a graph, with vertices V and edges E. We represent each edge $e \in E$ as an unordered pair $e = \{v, v'\}$, where $v, v' \in V$, thus e is an edge from v to v'. For convenience we'll label the vertices as $V = \{1, 2, \dots, n\}$. We write $|V| = n$ for the number of vertices of G, similarly $|E|$ for the number of edges. Given a vertex $v \in V$, the *neighbours* of V are vertices v' such that $\{v, v'\} \in E$. The *degree* $\deg(v)$ of vertex v is the number of neighbours of v.

A *stable set* S is a subset of V such that no two elements of S are neighbours (the term *independent* set is also often used for this). Here are some examples of stable sets, shown in red.



The question we want to ask is: how large can a stable set be?

Lemma 1.2 *Let $d = \frac{2|E|}{|V|}$ and suppose that $d \geq 1$. Then G contains a stable set of vertices with at least $\frac{|V|}{2d}$ elements.*

There's nothing explicitly connected to probability here, but you might sense that it will creep in by noticing that d is the average degree of vertices within G. To see this, note that $\frac{1}{n}\sum_v \deg(v) = \frac{1}{|V|}\sum_v \sum_e \mathbb{1}_{\{v \text{ is an endpoint of } e\}}$, and the summation will count each edge exactly twice, resulting in $\frac{2|E|}{|V|}$.

Proof of Lemma 1.2: Within the proof we'll write $|V| = n$ and $|E| = m$, to make the notation nicer.

Let $p \in (0, 1)$, to be chosen explicitly later. Let S be a random subset of V where for each $v \in V$ we insert v into S with probability p, independently for each v. Noting that $|S| = \sum_{v \in V} \mathbb{1}_{\{v \in S\}}$ we have

$$\mathbb{E}[|S|] = \sum_{v \in V} \mathbb{P}[v \in S] = np.$$

Let's now discuss strategy. The set S is (very probably) not a stable set. We need an extra step here that we didn't require in Problem 1.1 – we need to ask how close S is to being the object that we need to construct. That is, we ask how vertices we need to remove from S in order to make it into a stable set.

Let Y denote the set of edges $e \in E$ such that both endpoints of e are in S. If we take each edge $e \in Y$ and remove one of the endpoints of e from S (it does not matter which one) then the result will be a stable set of size at least $|S| - |Y|$. The 'at least' comes from the possibility that we might remove the same vertex as an endpoint of more than one $e \in Y$.

In symbols, $|Y| = \sum_{\{v,v'\} \in E} \mathbb{1}_{\{v \in S\}} \mathbb{1}_{\{v' \in S\}}$. Similarly to the calculation above, we have

$$\mathbb{E}[|Y|] = \sum_{\{v,v'\} \in E} \mathbb{P}[v \in S \text{ and } v' \in S] = \sum_{\{v,v'\} \in E} \mathbb{P}[v \in S]\mathbb{P}[v' \in S] = mp^2$$

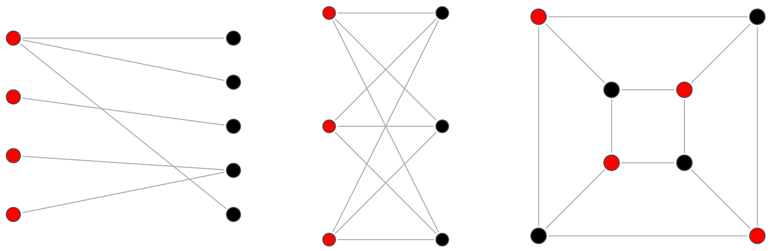
Note that we used the independence involved in the construction of S. We have assumed that $d = \frac{2m}{n}$, hence

$$\mathbb{E}[|S| - |Y|] = np(1 - \tfrac{1}{2}dp).$$

We now choose $p \in (0,1)$ to make this quantity as large possible. A bit of calculus finds the right choice, which is $p = \frac{1}{d}$, resulting in $\mathbb{E}[|S| - |Y|] = \frac{n}{2d}$. In particular, the probability that $|S| - |Y| \geq \frac{n}{2d}$ is strictly positive, which completes the proof. ■

With some harder work the bound in Lemma 1.2 can be improved from $\frac{|V|}{2d}$ up to $\frac{|V|}{d+1}$. This is known to be the best possible bound, for general graphs. It can also be proved using the probabilistic method, as well in other ways. As far as I know, in this case the probabilistic argument first appeared within the book of Alon and Spencer (2008). This book contains many examples of existence questions that can be solved using probabilistic constructions.

Exercise 1.3 Let $G = (V, E)$ be a graph. Here are two new pieces of terminology. We say that G is *bipartite* if we can divide the vertices of G into two disjoint sets $V = V_1 \cup V_2$ such that every $e \in E$ has precisely one endpoint in V_1 and the other endpoint in V_2. Here's an illustration of some bipartite graphs:



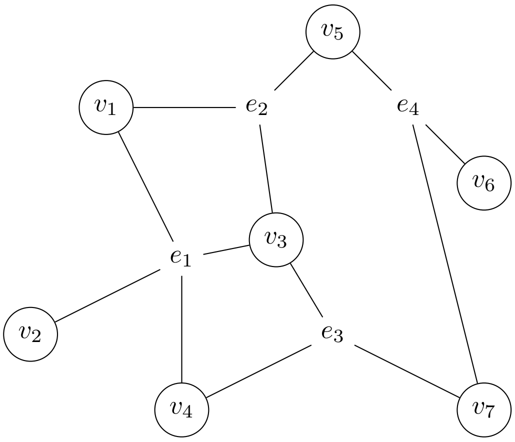
A *subgraph* $G' = (V', E')$ of G is a vertex set $V' \subseteq V$ with an edge set $E' \subseteq E$, such that each $e \in E'$ has both its endpoints in V'.

Let $G = (V, E)$ be a graph with m edges. Use the probabilistic method to show that G has a bipartite subgraph containing at least $\frac{m}{2}$ edges.

1.2 Hypergraph colouring

A *colouring* of a graph $G = (V, E)$ is an assignment of k colours to V, that is each vertex is given a colour. The colouring is said to be *proper* if, for each edge e, the vertices at either end of e are different colours. In this case the colouring is generally referred to as a *k-colouring* of G. We're going to think about the generalization of this idea to hypergraphs.

In a graph, each edge connects two vertices together, meaning that $e \in V^2$ as in the previous section. In a *hypergraph* a single edge connects together multiple vertices, so in this case we have $e \subseteq V$. Hence, E is a subset of the power set of V. In this context edges are sometimes known as hyperedges, but we'll stick with edge. There are various different ways to draw hypergraphs, for example:



We need to update our concepts slightly for this new situation. The *degree* $\deg(v)$ of the vertex $v \in V$ is the number of edges which contain that vertex. The *order* of an edge $|e|$ is the number of vertices contained within that edge. We say that a hypergraph is *k-regular* if each vertex has degree k, and *k-uniform* if each edge contains precisely k vertices.

A colouring of a hypergraph $G = (V, E)$ is an assignment of k colours to V; as before, each vertex is given a colour. The colouring is said to be *proper* if, for each edge e, the vertices within e are not all of the same colour. In this case we use the term k-colouring. There is a subtle point here: we do *not* require that for a given edge $e = \{v_1, \ldots, v_{|e|}\}$ all the v_i are different colours, just that they are not all the same colour. Note that when $|e| = 2$ this reduces to the definition for (not-hyper) graphs.

We'll restrict ourselves to finite hypergraphs, that is $|V| < \infty$. It is hopefully clear that if $k = |V|$ then the hypergraph $G = (V, E)$ has a k-colouring, where we simply take each vertex to be a different colour. The question is, can we find a k-colouring of G with fewer colours? There are lots of variations on this theme, but we'll pick on the following result.

Lemma 1.4 *Let $k \geq 9$ and let $G = (V, E)$ be a k-uniform, k-regular hypergraph. Then there exists a 2-colouring of G.*

Since we'll work with two colours from now on, let's agree to use red and black. We can see one idea of how probability might enter in using a similar method to that of Lemma 1.2: we could try colouring each vertex independently red/black with some probability $p \in (0, 1)$. However, in this case the probability of getting a 2-colouring is extremely small, and we wouldn't get anywhere near proving Lemma 1.4. We need a better idea, and it comes via the following lemma.

Lemma 1.5 (Lovász Local Lemma) *Let $G = (V, E)$ be a hypergraph. Let $p \in [0, 1]$ and $d \in N$. Let $A_1, \ldots, A_n$ be a sequence of events such that $\mathbb{P}[A_i] \le p$ for all i, and each event is independent of all except d of the others. If $\mathrm{e}p(d + 1) \le 1$ then $\mathbb{P}[A_1^c \cap A_2^c \cap \ldots \cap A_n^c] > 0$.*

Where to find it: Since its discovery by Erdős and Lovász (1975), plus some refinements thereafter, this result has become a well known tool in combinatorics and you will find it easily within textbooks. You might find it within a third year course on combinatorics or graph theory, but I expect this will depend on the personal taste of the lecturer. ■

Note that, in the statement of the lemma, the constant e is Euler's constant $\mathrm{e} \approx 2.72$, written without italics to distinguish it from an edge $e \in E$. We use A^c to denote the complement of the event A. The amazing thing about Lemma 1.5 is that it doesn't matter how many events $(A_i)_{i=1}^n$ we have, its still possible to avoid all of them happening, with positive probability.

There version we've stated here is more properly known as the symmetric Lovász local lemma. There is also an asymmetric version in which p may vary for different events. In both cases an explicit lower bound is known for $\mathbb{P}[A_1^c \cap \ldots \cap A_n^c]$, but we won't need that here. With Lemma 1.5 in hand, the proof of Lemma 1.5 is surprisingly easy, as follows.

Proof of Lemma 1.4: Colour each vertex $v \in V$, at random and independently of each other, either red or black with equal probability $\frac{1}{2}$. For each edge $e \in E$ let A_e be the event that all vertices within e are the same colour. Since G is k-regular we have $\mathbb{P}[A_e] = 2^{-(k-1)}$ for all $e \in E$.

Note that if $e, e' \in E$ are two edges and have no vertex in common, then A_e and $A_{e'}$ are independent. Any given edge $e \in E$ contains k vertices, and each such vertex can be contained in at most k edges (including e itself). Hence each event A_e is independent of all other $A_{e'}$ except for (at most) $k(k - 1)$ of them.

We therefore apply Lemma 1.5 with $d = k(k - 1)$ and $p = 2^{-(k-1)}$, which requires that $\mathrm{e}k(k - 1)2^{-(k-1)} \le 1$. It is straightforward to check that this inequality is satisfied when $k \ge 9$. Hence $\mathbb{P}[\bigcap_{e \in E} A_e^c] > 0$. On this event we have a 2-colouring of G, which therefore must exist. ■

To finish the story we might ask what happens when $k < 9$. It has been shown in Henning and Yeo (2013) that Lemma 1.4 holds when $k \ge 4$. Many variants of this question have been looked at in many different places but as far as I know, for this one, the cases $k = 2$ and $k = 3$ remain open.

Chapter 2

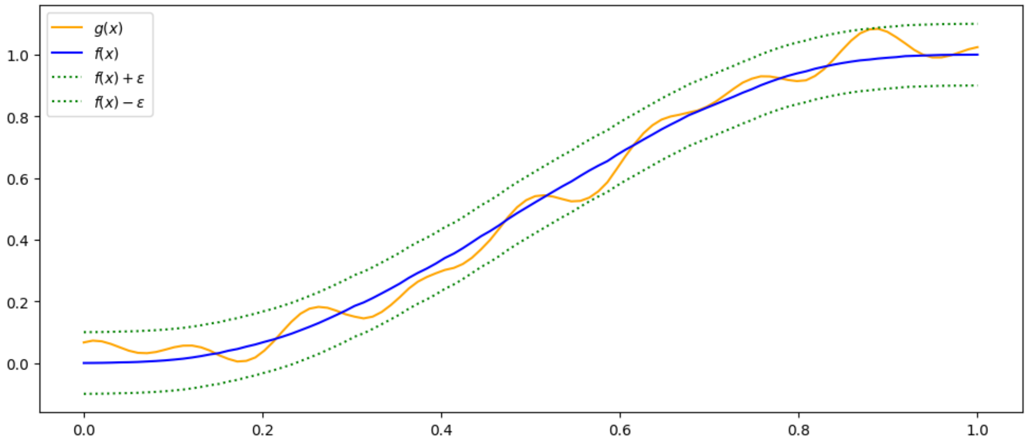
Polynomial approximation

2.1 The Weierstrauss approximation theorem

Weiersrauss' approximation theorem allows us to approximate continuous functions with polynomials. It is the basis of a useful technique for proving facts about continuous functions: first prove it for polynomials, then extend it from polynomials to all continuous functions. In order to carry out this extension, it is helpful to have polynomials that are very good approximations of continuous functions. Weiersrauss' theorem gives us that for any $\epsilon > 0$ and continuous function $f : [0,1] \to \mathbb{R}$, there exists a polynomial $g : [0,1] \to \mathbb{R}$ such that

$$\sup_{x\in[0,1]} |f(x) - g(x)| \le \epsilon. \tag{2.1}$$

In words, g can never stray more than ϵ away from f. It is perhaps best illustrated with a picture.



There is nothing special about $[0,1]$. Using a linear transformation we could move this result to any $[a,b] \subseteq \mathbb{R}$, but it is convenient to fix the endpoints as $[0,1]$.

Naturally, this has no obvious connection to probability. You'll usually find a proof based on real analysis in a third/fourth year course on functional analysis, but of course we don't want that one. We'll give a proof that relies on the weak law of large numbers. Let us now move in that direction, starting with a lemma that controls how close a random variable X is to its expected value $\mathbb{E}[X]$.

Lemma 2.1 (Chebyshev's Inequality) *Let X be a random variable and suppose that $\mathbb{E}[X^2] < \infty$. Then, for all $\epsilon > 0$ we have*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \frac{\operatorname{var}(X)}{\epsilon^2}. \tag{2.2}$$

Where to find it: This crops up in various different places. You might use it within first or second year probability, but for a proof you'll probably have to wait until third year, even though its quite an easy one. It is sometimes found inside courses on Lebesgue integration and/or measure theory. As you might imagine, it is closely related to Markov's inequality $\mathbb{P}[|X| \geq x] \leq \frac{1}{x}\mathbb{E}[|X|]$, where $x > 0$.

Several of the results in this course feature the condition $\mathbb{E}[X^2] < \infty$. This condition guarantees that $\mathbb{E}[X]$ and $\operatorname{var}(X)$ are both well defined, in the sense of Lebesgue integration. The connections between Lebesgue integration and probability are often covered in third year probability courses, occasionally in second year. There is no need to worry about this for our purposes. ■

Theorem 2.2 (Weak Law of Large Numbers) *Let $(X_i)_{i\in\mathbb{N}}$ be independent, identically distributed random variables, with mean $\mu = \mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2] < \infty$. Then for each $\epsilon > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right] \to 0.$$

The weak law of large numbers says something very natural: if you take lots of samples $(X_1, \ldots, X_n)$ of some random variable X, and then look at the mean average $\frac{1}{n}\sum_{i=1}^{n} X_i$ of your sample, this mean average will usually be close to $\mathbb{E}[X]$. It is called the 'weak' law because, with much more work, it is possible to prove that these two quantities become close in an even stronger sense, but we won't need that here. The basic idea is that some of the X_i will fall above μ and others will fall below but, because of independence, when you sum all of their values together there will lots of cancellation between the above and below parts.

Proof of Theorem 2.2: We plan to apply Lemma 2.1, so let us calculate the mean and variance of $\frac{1}{n}\sum_{i=1}^{n} X_i$. Clearly,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu.$$

Recall that $\operatorname{var}(X+Y) = \operatorname{var}(X) + 2\operatorname{cov}(X,Y) + \operatorname{var}(Y)$. Due to independence we have $\operatorname{cov}(X_i, X_j) = 0$ when $i \neq j$, hence

$$\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\operatorname{var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Therefore, Lemma 2.1 gives that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2}{\epsilon^2 n}.$$

This converges to zero as $n \to \infty$, as required. ■

Theorem 2.3 (Weierstrass' Approximation Theorem) *Let $f : [0,1] \to \mathbb{R}$ be continuous. Then, for any $\epsilon > 0$, there exists a polynomial $g : [0,1] \to \mathbb{R}$ such that*

$$\sup_{x\in[0,1]} |f(x) - g(x)| \leq \epsilon.$$

Proof: Let $x \in [0,1]$ and let (X_i) be independent Bernoulli trials with success probability x, that is $\mathbb{P}[X_i = 1] = x$ and $\mathbb{P}[X_i = 0] = 1 - x$. Write $S_n = \sum_{i=1}^n X_i$, which has the Binomial(n, x) distribution, and define the polynomial

$$g(x) = \mathbb{E}\left[f\left(\tfrac{S_n}{n}\right)\right] = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f(\tfrac{k}{n}). \tag{2.3}$$

Our plan is to use the weak law of large numbers to compare g to f.

Let $\epsilon > 0$. Recall (from real analysis) that continuous functions on closed bounded intervals are uniformly continuous. Hence there exists $\delta > 0$ such that $|f(x) - f(y)| \leq \epsilon$ for all $x, y \in [0,1]$ that satisfy $|x - y| \leq \delta$. We have

$$\begin{aligned}
|g(x) - f(x)| &= \mathbb{E}\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right|\right] \\
&= \mathbb{E}\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right| \,;\, \left|\tfrac{S_n}{n} - x\right| > \delta\right] \mathbb{P}\left[\left|\tfrac{S_n}{n} - x\right| > \delta\right] \\
&\qquad + \mathbb{E}\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right| \,;\, \left|\tfrac{S_n}{n} - x\right| > \delta\right] \mathbb{P}\left[\left|\tfrac{S_n}{n} - x\right| > \delta\right] \\
&\leq \epsilon \times 1 + 2||f||_\infty \times \mathbb{P}\left[\left|\tfrac{S_n}{n} - x\right| > \delta\right].
\end{aligned} \tag{2.4}$$

Here, to deduce the second line we partition on the event $\{|\frac{S_n}{n} - x| > \delta\}$. Note that we are writing conditional expectation as $\mathbb{E}[X\,;\,A]$ instead of $\mathbb{E}[X \mid A]$, to avoid confusion with modulus signs. To deduce the third line, in the first term we use the uniform continuity property from above and the fact that probabilities are bounded above by one. For the second term, we use that $|f(\frac{S_n}{n}) - f(x)| \leq |f(\frac{S_n}{n})| + |f(x)| \leq 2||f||_\infty$ where $||f||_\infty = \sup\{|f(x)|\,;\, x \in [0,1]\}$.

Letting $n \to \infty$ in (2.4), Theorem 2.2 gives that the second term tends to zero, leaving us with $|g(x) - f(x)| \leq \epsilon$. We have shown this for any $x \in [0,1]$ so this completes the proof. ■

The function defined in (2.3) is known as a Bernstein polynomial. Note that we have done a bit better in the proof than we had claimed in the theorem: not only does a polynomial g with the desired property exist, we can also write g down explicitly. For this reason the probabilistic proof of Theorem 2.3 is quite well known. The version above was adapted from McKean (2014), but apart from differences in notation it is much the same as Bernstein's original proof from a century earlier.

The type of argument that led to (2.4) is quite common in probability theory. We knew that the event $\{|\frac{S_n}{n} - x| > \delta\}$ was unlikely, which allowed us to get rid of it and concentrate on the likely event $\{|\frac{S_n}{n} - x| \leq \delta\}$, and in *that* case we could deduce what we wanted. This is a bit different to the probabilistic method of Chapter 1. Here we constructed a deterministic object using probability, and used probability to obtain information about its properties.

Exercise 2.4 (Bezier curves and computer graphics) We used $f, g : [0,1] \to \mathbb{R}$ in Theorem 2.3 but in fact it holds for $f, g : [0,1] \to \mathbb{R}^d$. Here g becomes a polynomial function $g : [0,1] \to \mathbb{R}^d$ which has the form $g(x) = \sum_{i=0}^n \mathbf{c}_i x^i$ where $\mathbf{c}_i \in \mathbb{R}^d$. The proof is much the same.

Let $(\mathbf{P}_i)_{i=1}^n$ be a finite sequence of points in $\mathbb{R}^d$. The *Bezier curve* of $(\mathbf{c}_i)$ is the Bernstein polynomial $B_{\mathbf{P}_0 \dots \mathbf{P}_n}(t)$ defined by

$$t \mapsto \sum_{i=1}^{n} \binom{n}{i} t^i (1-t)^{n-i} \mathbf{P}_i$$

for $t \in [0,1]$. Note that $B_{\mathbf{P}_0 \dots \mathbf{P}_n}(0) = \mathbf{P}_0$ and $B_{\mathbf{P}_0 \dots \mathbf{P}_n}(1) = \mathbf{P}_n$ and that if $n = 2$ then $B_{\mathbf{P}_0 \dots \mathbf{P}_n}$ is simply a straight line segment from $\mathbf{P}_0$ to $\mathbf{P}_1$.

(a) Show that $B_{\mathbf{P}_0 \dots \mathbf{P}_n}(t) = tB_{\mathbf{P}_0 \dots \mathbf{P}_{n-1}}(t) + (1-t)B_{\mathbf{P}_1 \dots \mathbf{P}_n}(t)$.

(b) Consider the case $n = 2$. Use part (a) to explain why the following algorithm draws the Bezier curve $B_{\mathbf{P}_0 \dots \mathbf{P}_2}$.

```
For each $t \in [0,1]$:
  1. Make a straight line starting from $\mathbf{P}_0$ and ending at $\mathbf{P}_1$. Let $\mathbf{Q}_0$ be the point that is a
     fraction $t$ of the way along this line.
  2. Make a straight line starting from $\mathbf{P}_1$ and ending at $\mathbf{P}_2$. Let $\mathbf{Q}_1$ be the point that is a
     fraction $t$ of the way along this line.
  3. Make a straight line starting from $\mathbf{Q}_0$ and ending at $\mathbf{Q}_1$. Draw the point that is a fraction
     $t$ of the way along this line.
```

(c) In (b), explain briefly why for $t \approx 0$ the direction of $B(t)$ is approximately the same as that of the line segment from $\mathbf{P}_0$ to $\mathbf{P}_1$, and why for $t \approx 1$ the direction of $B(t)$ is approximately the same as that of the line segment from $\mathbf{P}_1$ to $\mathbf{P}_2$,

(d) From what you've learned in (b), can you now explain a recursive algorithm based on (a) to draw the curve $B_{\mathbf{P}_0 \dots \mathbf{P}_n}$?

(e) What does Theorem 2.3 tell you about this situation?

The point of the above exercise is computer graphics and animation. Computers store shapes as ordered lists of coordinates, so the curve representing (for example) the outline of an arm would simply be a list $\mathbf{P}_0 \dots \mathbf{P}_n$. It is tempting to say: draw an arm by connecting each pair $\mathbf{P}_k\mathbf{P}_{k+1}$ with a straight line segment. Storing a long enough list to make the result of this process actually look like an arm on a cinema sized screen would require hundreds, if not thousands, of points. Generating animated segments of film that way is computationally infeasible.

Here's a better idea. Use a shorter list $\mathbf{P}_0 \dots \mathbf{P}_n$ and do the following. Take the points $\mathbf{P}_0 \dots \mathbf{P}_n$ and consider successive triplets

$$(\mathbf{P}_0 \dots \mathbf{P}_2),\ (\mathbf{P}_2 \dots \mathbf{P}_4),\ (\mathbf{P}_4 \dots \mathbf{P}_6),\ \dots,\ (\mathbf{P}_{n-2} \dots \mathbf{P}_n). \tag{2.5}$$

Then, pick some $\epsilon > 0$ and for each triplet $\mathbf{P}_k \dots \mathbf{P}_{k+2}$ use the algorithm in (b) to draw the points $B_{\mathbf{P}_k \dots \mathbf{P}_{k+2}}(k\epsilon)$ for $k\epsilon \in [0,1]$. Take $\epsilon > 0$ small enough that, when you connect *these* points together with straight line segments, thanks to part (c) it looks smooth, and more like an arm.

You should ask: why not simply calculate and then draw points on $B_{\mathbf{P}_0 \dots \mathbf{P}_n}$, and connect them together with straight lines? This does a good job, but for visual purposes it isn't noticeably better that the above, and it uses a lot more computation. Sometimes quadruplets of points are used in place of (2.5), in which case the $n = 3$ case of the algorithm in (d) is needed to paint in the details; this require more calculations but gives a smoother output.

That is how computer animation works. A moving 'skeleton' of points $\mathbf{P}_0 \dots \mathbf{P}_n$ defines an object and the algorithm above paints in the details, frame by frame. For still images it is usually referred to as vector graphics, which is used for clipart libraries and fonts, including the letters on this page. They key point is that the algorithm we've just described only relies on computations involving straight lines. Modern graphics cards are *really* fast at computations that only involve straight lines (i.e. linear functions) and the algorithm we've outlined suits them perfectly. This allows computers to simulate very complex natural looking scenes, as you already know from watching films and playing games.

2.2 Taylor's theorem -ish

Theorem 2.3 is useful, but suppose that you want a polynomial approximation that is exact at some chosen point x_0, and a reasonable approximation nearby. Then, what you want is a Taylor series. I'm sure you all know Taylor's theorem well enough that I don't need to write it down, but here's a thought: what if you wanted to use Taylor's theorem to approximate a function $f : \mathbb{R} \to \mathbb{R}$ near $f(x_0) = a$, with a polynomial, but f was not differentiable? Then $f'(x_0)$, $f''(x_0)$ and so on, don't exist, and you can't even write down a Taylor expansion, never mind worry about whether it converges. This *is* a bit of a frivolous situation. You won't want to do this often. Perhaps surprisingly, it can still be done.

In fact it can be done using a variation of the techniques that we used to prove Weierstrauss' approximation theorem. We'll need the *discrete derivative*

$$[\Delta_h f](x) = \frac{f(x+h) - f(x)}{h}.$$

Applying the Δ_h operator m times, we get the m^{th} order discrete derivative, given by

$$[\Delta_h^{(m)} f](x) = \frac{1}{h^m} \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} f(x + jh).$$

If f is differentiable m times at x then it can be checked that $[\Delta_h^{(m)} f](x) \to f^{(m)}(x)$ as $h \to 0$. Note that $[\Delta_h^{(m)} f](x)$ still makes sense even if f is not differentiable.

The following result comes from numerical analysis. It is mostly used in cases where f is differentiable, but the computer doesn't need to know about that. Having said that, as we will see in the next chapter, *probability theory* involves a lot of non-differentiable functions. Once again, there is no need to restrict to $[0, 1]$, any closed bounded interval will do.

Theorem 2.5 (Newton-Hille Interpolation) *Let $f : [0, 1] \to \mathbb{R}$ be continuous and let $x \in [0, 1]$. Then for all y such that $x + y \in [0, 1]$, as $h \downarrow 0$ we have*

$$\sum_{m=0}^{\infty} \frac{y^m}{m!} [\Delta_h^{(m)} f](x) \to f(x + y).$$

Note that replacing $(\Delta_h^{(m)} f)$ by the m^{th} derivative $f^{(m)}$ results in the infinitely differentiable case of Taylor's formula (and h then does not appear). More formally, we can recover a version of Taylor's theorem from Theorem 2.5 by adding some extra conditions that allow us to take a term-by-term limit of the summation, as $h \to 0$, but we won't do that here. Note also that Theorem 2.5 only applies with one sided limits at the endpoints of $[0, 1]$.

Proof of Theorem 2.5: Without loss of generality we will take $x = 0$. The result for general $x \in [0,1]$ can be recovered by translation $a \mapsto a + x$ and (for $y < 0$) reflection. We will also restrict our proof to the subsequence $h = \frac{1}{n}$ as $n \to \infty$.

Let $(X_i)_{i\in\mathbb{N}}$ be independent, identically distributed random variables, each with the Poisson(λ) distribution where $\lambda = y$. We prefer to write λ as this is the usual notation for the Poisson parameter. For independent Poisson random variables, it holds that Poisson(a) + Poisson(b) has the same distribution as Poisson($a + b$). This might appear when you study tranformations of random variables in first/second year, but it will certainly appear the first time you study the Poisson process in second/third year. It follows that $S_n = \sum_{i=1}^n X_i$ has the Poisson($n\lambda$) distribution.

We'll now use the same strategy as in the proof of Theorem 2.3: we will apply the weak law of large numbers to $\mathbb{E}[f(\frac{S_n}{n})]$ and also calculate it explicitly. For the explicit calculation we have

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] &= \sum_{i=0}^\infty e^{-\lambda n}\frac{(\lambda n)^i}{i!}f(\tfrac{i}{n})\\ &= \sum_{i=0}^\infty\sum_{j=0}^\infty \frac{(-\lambda n)^j}{j!}\frac{(\lambda n)^i}{i!}f(\tfrac{i}{n})\\ &= \sum_{i=0}^\infty\sum_{j=0}^\infty \frac{(-1)^j}{i!j!}\left(\frac{\lambda}{h}\right)^{i+j} f(ih) \qquad \text{writing } h = \frac{1}{n}\\ &= \sum_{m=0}^\infty \left(\frac{\lambda}{h}\right)^m \frac{1}{m!}\sum_{i=0}^m \binom{m}{i}(-1)^{m-i}f(ih) \qquad \text{writing } m = i + j\end{aligned}$$

Here, to deduce the final line we have noted that $\frac{1}{i!j!} = \frac{m!}{i!(m-i)!}$ follows from $m = i+j$, and we have exchanged the order of summation. Note that it is straightforward to check absolute convergence by comparison to $(e^{\lambda n})^2$, along with the fact that continuity of f implies $\sup\{|f(x)|\,;\, x \in [0,1]\} < \infty$. Writing the above in the notation of discrete derivatives, we obtain

$$\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] = \sum_{m=0}^\infty \frac{\lambda^m}{m!}[\Delta_h^{(m)} f](0).$$

The same calculation as led to (2.4) also applies here, except that we should replace $x = \mathbb{E}[X_i]$ from (2.4) with our present $\mathbb{E}[X_i] = \lambda$. Again using Theorem 2.2, we reach the conclusion that $\mathbb{E}[f(\frac{S_n}{n})] \to f(\lambda)$ as $n \to \infty$, as required. ■

The proofs of Theorems 2.3 and 2.5 were surprisingly similar. We might ask whether putting in other common distributions into Theorem 2.2, and trying a similar proof, will give anything interesting. The answer is yes – using the gamma distribution leads to a non-standard inversion formula for the Laplace transform, and using the beta distribution leads to solving a version of the *moment problem* (where a random variable is asked to be constructed with specified values for the $\mathbb{E}[X^n]$). There is obviously something deep going on here, but I don't know what it is. You can find the calculations in Goldstein (1975).

2.3 Infinitely differentiable functions

Let I be an open subset of $\mathbb{R}$. If $f : I \to \mathbb{R}$ can be differentiated infinitely many times, at all points $x \in I$, then we say that f is *infinitely differentiable* on I. This concept is closely related to Taylor series, but there are some subtleties in the connection that we will explore in this section. If an infinitely differentiable function f satisfies

$$f(x) = \sum_{n=0}^{\infty} \frac{(x-x_0)^n}{n!} f^{(n)}(x_0) \tag{2.6}$$

for all $x, x_0 \in I$ then we say that f is *analytic* in I. Equation (2.6) says that within I the function f is always equal to its own Taylor series.

You might hope that all infinitely differentiable functions would be analytic, but this is not the case. For example,

$$f(x) = \begin{cases} 0 & \text{for } x \le 0, \\ e^{-1/x} & \text{for } x > 0, \end{cases} \tag{2.7}$$

satisfies $f^{(n)}(x) = 0$ for all $x < 0$ and

$$f^{(1)}(x) = \frac{1}{x^2} e^{-1/x}$$
$$f^{(2)}(x) = \left(\frac{1}{x^2} - \frac{2}{x^3}\right) e^{-1/x}$$

and so on for all $x > 0$. Using that $\frac{e^{-1/x}}{x^m} \to 0$ as $x \to 0$, it is straightforward to show that the left and right n^{th} derivatives at 0 are equal to 0 for all n. Hence f is infinitely differentiable, but the Taylor series for f at 0 is simply $\sum_{n=0}^{\infty} \frac{x^n}{n!} 0 = 0$ which doesn't match $f(x)$ for $x > 0$. You might come across this example in integration courses, for various reasons.

In the above example, if we looked at the Taylor series about some point $x_0 \neq 0$, then we would discover that the resulting Taylor series did match f within some small open interval $(x_0 - \epsilon, x_0 + \epsilon)$. We could still use Taylor series for (2.7), just not at $x = 0$. However, we might wonder how bad things can get. We say that a function $f : I \to \mathbb{R}$ is *nowhere analytic* if f is not analytic on any non-empty open interval $O \subseteq I$. For these functions, we can't use a Taylor series anywhere.

Theorem 2.6 (Fabius' Function) *There exists a function $f : \mathbb{R} \to \mathbb{R}$ that is infinitely differentiable and nowhere analytic.*

In fact, lots of examples of infinitely differentiable and nowhere analytic functions are known to exist. We will construct one such f explicitly, following a famous paper by Fabius (1966). It will take quite a bit of work. The strategy is to construct a function f that satisfies a particular differential equation,

$$f'(x) = 2f(2x) \tag{2.8}$$

and to show that any non-zero solution of this equation has the required properties. Note that the existence of a solution to (2.8) is not covered by the usual existence theorems for ODEs (e.g. Picard's theorem) because of the $f(2x)$ term. In fact solutions to (2.8) are non-unique; $f(x) = 0$ is one solution, and we will construct a non-zero solution with $f(0) = 0$.

The role of probability here is to provide the non-zero solution. Let's first establish something in the right direction and then we'll think about the rest of the proof. It be helpful to use a convention about one sided derivatives. When say that a differential equation such as (2.8) holds for $x \in [a, b]$, we mean with two-sided derivatives at $x \in (a, b)$ and one-sided derivatives at a and b.

Lemma 2.7 *Let (U_i) be a sequence of i.i.d. uniform random variables on $[0,1]$. Set*

$$X = \sum_{n=1}^{\infty} \frac{U_n}{2^n}$$

and let $F(x) = \mathbb{P}[X \le x]$ for $x \in \mathbb{R}$. Then

- *for $x \in [0, \frac{1}{2}]$ we have $F'(x) = 2F(2x)$,*
- *for $x \in [\frac{1}{2}, 1]$ we have $F'(x) = -2F(2x-1) + 2$.*

Proof: Note that $\sum_{n=1}^{\infty} 2^{-n} = 1$, so it is clear that $\mathbb{P}[0 \le X \le 1] = 1$, hence $F(0) = 0$ and $F(1) = 1$. Hence also $F(x) = 0$ for all $x \le 0$ and $F(x) = 1$ for all $x \ge 1$. Let (U_n) and X be as in the statement of the lemma and define also

$$X' = 2\sum_{n=2}^{\infty} \frac{U_n}{2^n}. \tag{2.9}$$

Clearly $X' = \sum_{n=1}^{\infty} \frac{U_{n+1}}{2^n}$, so X' and X have the same distribution. From (2.9) we have that $X' = 2(X - \frac{1}{2}U_1) = 2X - U_1$, hence $X = \frac{1}{2}(X' + U_1)$. Note also that X' and U_1 are independent. Combining these facts, X and $\frac{1}{2}(X+U)$ have the same distribution, where $U \sim \text{Uniform}([0,1])$ independently of X. Hence

$$\begin{aligned}\mathbb{P}[X \le x] &= \mathbb{P}[\tfrac{1}{2}(X + U_1) \le x] \\ &= \mathbb{P}[X \le 2x - U_1] \\ &= \int_0^1 \mathbb{P}[X \le 2x - u]\, du. \end{aligned} \tag{2.10}$$

The first line follows from our discussion above and the second is elementary. The third line might use slightly more than you are used to seeing, so let us briefly explain. If X had probability density function $f_X(x)$ then we could note that, by independence, the joint p.d.f. of (X,U) is $f_X(x)f_U(u)$, which would give $\mathbb{P}[X \le 2x + U_1] = \int_0^1 \int_0^{2x-u} f_X(x) f_U(u)\, dx\, du = \int_0^1 \mathbb{P}[X \le 2x - u] f_U(u)\, du$, and then since $f_U(u) = 1$ for $u \in [0,1]$ we would have the third line. However, we don't know that X has a probability density function (in fact, it doesn't). Without a p.d.f. we require a bit of machinery from Lebesgue integration, which you'll see if you take a course on it, but the result comes out as the same here. The key idea is that X and U are independent, so we can make use of the p.d.f. of U without having to touch X. Writing (2.10) in terms of F, we have

$$F(x) = \int_0^1 F(2x-u)\, du = \int_{2x-1}^{2x} F(t)\, dt. \tag{2.11}$$

The second equality here is deduced via the substitution $t = 2x - u$. We must now split into two cases.

Let us first consider $x \in [0, \frac{1}{2}]$. For such x we have $2x - 1 \le 0 \le 2x \le 1$. Noting that $F(t) = 0$ for $t \le 0$ we obtain from (2.11) that $F(x) = \int_0^{2x} F(t)\, dt$. Differentiating, which requires the fundamental theorem of calculus, we obtain that $F'(x) = 2F(2x)$.

It remains to consider $x \in (\frac{1}{2}, 1]$. For such x we have $0 \le 2x - 1 \le 1 \le 2x \le 2$. Noting that $F(t) = 1$ for $t \in [1,2]$ we obtain from (2.11) that $F(x) = \int_{2x-1}^{1} F(t)\, dt + 2x - 1$. Differentiating obtains that $F'(x) = -2F(2x-1) + 2$. ■

Proof of Theorem 2.6: The function $F(x)$ defined in Lemma 2.7 has the property that we want for $F \in [0, \frac{1}{2}]$, but not outside of that. We will define a function $f(x)$ that is equal to $F(x)$ for $x \in [0,1]$, but is defined differently elsewhere, in such a way that $f(x) = 2f(2x)$ for all $x \in \mathbb{R}$. We will define f in three stages. There is a picture below the proof to illustrate each stage.

Stage 1: For $x \in [0,2]$. We define

$$f(x) = F(x) \qquad \text{for } x \in [0,1], \tag{2.12}$$

$$f(x) = 1 - f(x-1) \qquad \text{for } x \in [1,2]. \tag{2.13}$$

Note that the second line of the above makes use of the first, and that $F(1) = 1$ and $F(0) = 0$, so f is well defined at the joining point $x = 1$. Note that $f(2) = 1 - f(1) = 0$, in fact the extension in (2.13) is like placing a vertical mirror on the graph at $x = 1$.

Let $x \in [\frac{1}{2}, 1]$, which means that $0 \le 2x - 1 \le 1 \le 2x \le 2$. We therefore have

$$\begin{aligned} f'(x) &= F'(x) \\ &= -2F(2x-1) + 2 \\ &= -2f(2x-1) + 2 \\ &= -2(1 - f(2x)) + 2 \\ &= 2f(2x). \end{aligned}$$

The first line of the above uses (2.12). The second line uses Lemma 2.7, the third line uses (2.12) again, and the fourth line uses (2.13). We have now defined $f(x)$ for $x \in [0,2]$ and showed that (2.8) holds for $x \in [0,1]$.

Stage 2: Induction on $x \in [2^n, 2^{n+1}]$. We will use induction to simultaneously extend f and show that it satisfies (2.8). Note that, at the point where we have defined $f(x)$ for $x \in [0, A]$, we can make sense of (2.8) for $x \in [0, A/2]$, but no further than that. We will work with the following inductive hypothesis:

$(\mathrm{IH})_n$: The function $f(x)$ has been defined for $x \in [0, 2^n]$ and:

(a) for $x \in [0, 2^{n-1}]$ it satisfies equation (2.8),

(b) for $x \in [2^{n-1}, 2^n]$ it satisfies $f(x) = -f(x - 2^{n-1})$.

(c) for $x \in \{0, 2, 4, \dots, 2^n\}$ it satisfies $f(x) = 0$.

The base case is $(\mathrm{IH})_1$, which we have already proved in Stage 1.

Suppose that $(\mathrm{IH})_n$ holds. Thus $f(x)$ is defined for $x \in [0, 2^n]$, and we extend the definition by setting

$$f(x) = -f(x - 2^n) \qquad \text{for } x \in [2^n, 2^{n+1}]. \tag{2.14}$$

Note that for even integers $2^n + 2k$ between 2^n and 2^{n+1} we have $f(2^n + 2k) = f(2k) = 0$, by (c) of $(\mathrm{IH})_n$, which gives (c) of $(\mathrm{IH})_{n+1}$. Also, since $f(2^n) = 0$ we have that this extension is well defined at the joining point $x = 2^n$. Thus (2.14) gives part (b) of $(\mathrm{IH})_{n+1}$.

Let $x \in [2^{n-1}, 2^n]$, for which we have $0 \le x - 2^{n-1} \le 2^{n-1} \le 2x \le 2^n$ and hence

$$\begin{aligned} f'(x) &= -f'(x - 2^{n-1}) \\ &= -2f(2(x - 2^{n-1})) \end{aligned}$$

$$
\begin{aligned}
&= -2f(2x - 2^n) \\
&= -2(-f(2x)) \\
&= 2f(2x).
\end{aligned}
$$

Here, the first line uses (b) from $(\mathrm{IH})_n$, the second line uses (a) from $(\mathrm{IH})_n$, the third line is elementary and the fourth line uses (2.14). We therefore have part (a) of $(\mathrm{IH})_{n+1}$, which means we have established the whole of $(\mathrm{IH})_{n+1}$.

Stage 3: For $x \leq 0$. From stage 2 we have $f(x)$ defined and satisfying (2.8) for all $x \in [0, \infty)$. We finally define

$$f(x) = f(-x) \qquad \text{for } x \in (-\infty, 0]. \tag{2.15}$$

Note that $f(0) = 0$ so this is well defined at the joining point $x = 0$. With this definition, for $x \leq 0$ we have $f'(x) = -f'(-x) = -2f'(-2x) = -2f'(2x)$ so here we have a slightly different equation to (2.8),

$$f'(x) = -2f(2x) \qquad \text{for } x \in (-\infty, 0]. \tag{2.16}$$

Stage 4: We've now shown that (2.8) holds for all $x \in [0, \infty)$ and (2.16) holds for all $x \in (-\infty, 0]$. Let's concentrate on $x \in [0, \infty)$ first. Differentiating (2.8) n times, we have

$$f^{(n)}(x) = 2^{1+2+\ldots+n} f(2^n x) = 2^{\frac{1}{2}n(n+1)} f(2^n x). \tag{2.17}$$

In particular, f is infinitely differentiable at all $x \in [0, \infty)$. From Stage 2 we have $f(2k) = 0$ for all $k \in \mathbb{N}$, which means that if $x \in \mathbb{R}$ has the form

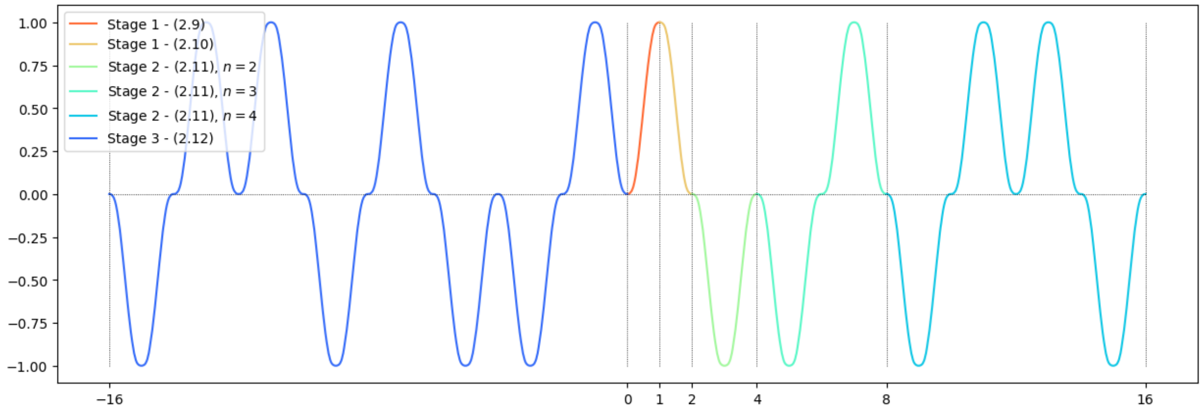
$$x = \frac{2j}{2^m} = \frac{j}{2^{m-1}} \qquad \text{for } m \in \mathbb{N}, j \in \mathbb{Z} \tag{2.18}$$

then $f^{(n)}(x) = 2^{\frac{1}{2}n(n+1)} f(2j2^{n-m}) = 0$ for all $n \geq m$. The same argument applies at $x \in (-\infty, 0]$, using (2.16) in place of (2.8), which just changes the $2^{\frac{1}{2}n(n+1)}$ in (2.17) to a $(-2)^{\frac{1}{2}n(n+1)}$. Strictly, we should check that all the left and right n^{th} derivatives at zero match up, but this is easy because they are all equal to zero.

Points of the form (2.18) are known as binary rationals. Any open interval contains infinitely many of them[1]. We have shown that if x is a binary rational then $f^{(n)}(x) = 0$ for all except finitely many $n \in \mathbb{N}$. Therefore, at such points, the Taylor expansion of f has only a finite number of non-zero terms, which makes it a polynomial. The function f is certainly not a polynomial (if it was, we would have $|f(x)| \to \infty$ as $x \to \infty$, and this contradicts $f(2k) = 0$) hence f is not analytic in any open interval of $\mathbb{R}$. ■

It's helpful to see what's going on in the above proof with a graph of f. We've already drawn it for $x \in [0, 1]$. In fact, it was the blue line in the picture illustrating (2.1) in Section 2.1! I doubt you noticed anything unusual about that picture, which should serve as a warning that Taylor series might fail to work when you least expect it. The extensions look like this:

[1] More formally, the binary rationals are *dense* in $\mathbb{R}$.



Stage 1 is the reflection of $\{f(x)\,;\,x\in[0,1]\}$ about the vertical line at $x=1$, to define $f(x)$ for $x\in[1,2]$. Stage 2, shown for $n=2,3,4$, flips whatever has already been defined for $[0,2^n]$ upside down, and then shifts it rightwards into $[2^n,2^{n+1}]$. Lastly, stage 3 is a reflection about the vertical line at $x=0$.

The function f can also be defined as the Fourier transform of $\prod_{m=1}^{\infty}\left(\cos\frac{\pi x}{2^m}\right)^m$. It was used by Jessen and Wintner (1935) to study the Riemann zeta function. This also involved probability, based on a connection between almost periodic functions and sums of independent random variables, but I don't know anything more than that.

The existence of non-analytic infinitely differentiable functions is very important in functional analysis and geometry as so-called 'test' functions. A test function is zero outside of some bounded set and is infinitely differentiable. For example take $T(x)=f(x)\mathbb{1}_{x\in[0,1]}$, where $T:\mathbb{R}\to\mathbb{R}$. They are used (in combination with integration) as a tool for focusing on events within a particular region of space.

Chapter 3

Counting prime divisors

Let $\nu(n)$ denote the number of primes p dividing $n \in \mathbb{N}$. This is meant entirely literally: we do *not* count with multiplicity, so for example $\nu(3^2) = \nu(3) = 1$. The other way might seem more elegant (and I don't know) but it is the function ν that was famously studied by Hardy and Ramanujan (1917). Their result concerns how large $\nu(n)$ becomes when $n \in \mathbb{N}$ is a 'typical' large number. We'll study a slight extension of Hardy and Ramanujan's result, which comes from Turán (1934).

We need to introduce a small piece of notation. If $(a_n)_{n\in\mathbb{N}}$ is a sequence then we write:

- $\mathcal{O}(a_n)$ in place of a sequence (b_n), provided that for some constant $C < \infty$ and all $n \in \mathbb{N}$ we have $|a_n| \leq C|b_n|$.

This is known as *big-O* or *Landau notation.* The point is to save time. We can write e.g. $n+\frac{\sqrt{e}\pi^7}{n}$ as simply $n+\mathcal{O}(\frac{1}{n})$ without needing to know exactly what denominator sits on top of the fraction (in this case $\sqrt{e}\pi^7$). Since the $\mathcal{O}(\frac{1}{n})$ bit will tend to zero we often don't want to know either. Note that $\mathcal{O}(1)$ simply denotes a bounded sequence. The result we are interested in within this chapter is the following.

Theorem 3.1 (Turán's Theorem, after Hardy-Ramanujan) *Let $w(n)$ satisfy $w(n) \to \infty$ as $n \to \infty$. Let $d(n)$ be the number of $x \in \{1, 2, \dots, n\}$ for which*

$$|\nu(x) - \log\log n| > w(n)\sqrt{\log\log n}.$$

Then $\frac{d(n)}{n} \to 0$ as $n \to \infty$.

In words, when n is large, nearly all $x \in \{1, \dots, n\}$ have approximately $\log\log n$ prime factors. That is not very many prime factors. There are a little over 8 billion people alive and $\log\log(8 \text{ billion})$ is slightly more than 3.

Before we start the proof of Theorem 3.1, we need a convention about sums over prime numbers. In this chapter, whenever we write a summation over the symbol p and/or q, we mean to sum only over the prime numbers. For example, $\sum_{p\leq n} \frac{1}{p}$ means to sum $\frac{1}{p}$ over all primes p such that $p \leq n$. On that matter we have the following lemma.

Lemma 3.2 *It holds that $\sum_{p\leq n} \frac{1}{p} = \log\log n + \mathcal{O}(1)$.*

Where to find it: You'll find this within third/fourth year number theory. It can be proved directly via Stirling's formula and Abel summation, or in prettier ways using Euler's product formula, but I'm no expert here. ■

Proof of Theorem 3.1: The connection to probability is (I hope) still invisible, but let us now make it. Let X be chosen uniformly at random from the set $\{1,\dots,n\}$. The probability that X is divisible by p is approximately $\frac{1}{p}$, which won't surprise you, and we'll make it precise below. The really key fact, which will come after, is that the events $\{X \text{ is divisible by } p\}$ and $\{X \text{ is divisible by } q\}$ are approximately independent, for distinct primes p and q.

For primes p set

$$X_p = \begin{cases} 1 & \text{if } p \text{ divides } X \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\nu(X) = \sum_{p\le n} X_p$. We will split the proof into two parts, the first of which is to investigate the X_p. In the second part we use the X_p to investigate $\nu(X)$.

Part 1: Recall that for $x\in\mathbb{R}$ the floor function $\lfloor x\rfloor$ denotes x rounded down to the nearest integer. There are exactly $\lfloor n/p\rfloor$ numbers in $\{1,\dots,n\}$ that are divisible by p, hence

$$\begin{aligned}\mathbb{E}[X_p] &= 1\times\mathbb{P}[p \text{ divides } X] + 0\times\mathbb{P}[p \text{ does not divide } X]\\ &= \mathbb{P}[p \text{ divides } X]\\ &= \frac{\lfloor n/p\rfloor}{n}\\ &= 1/p + \mathcal{O}(1/n).\end{aligned} \tag{3.1}$$

To deduce the last line of the above we use that $x-1\le\lfloor x\rfloor\le x$ for all $x\in\mathbb{R}$. In similar style, if p and q are distinct primes then

$$\begin{aligned}\mathbb{E}[X_pX_q] &= \mathbb{P}[p \text{ divides } X \text{ and } q \text{ divides } X]\\ &= \frac{\lfloor n/(pq)\rfloor}{n}\\ &= 1/(pq) + \mathcal{O}(1/n).\end{aligned} \tag{3.2}$$

Combining (3.1) and (3.2),

$$\mathbb{P}[p \text{ divides } X]\mathbb{P}[q \text{ divides } X] = \mathbb{P}[p \text{ divides } X \text{ and } q \text{ divides } X] + \mathcal{O}(1/n), \tag{3.3}$$

which is the 'approximate independence' mentioned at the start of the proof. More precisely, we will require upper bounds for $\mathrm{var}(X_p)$ and $\mathrm{cov}(X_p,X_q)$, where p and q are distinct primes. Noting that X_p is equal to either 0 or 1, we have $X_p = X_p^2$ and hence

$$\mathrm{var}(X_p) \;=\; \mathbb{E}[X_p^2] - \mathbb{E}[X_p]^2 \;=\; \mathbb{E}[X_p] - \mathbb{E}[X_p]^2 \;\le\; \mathbb{E}[X_p] \;=\; \tfrac{1}{p} + \mathcal{O}(\tfrac{1}{n}). \tag{3.4}$$

We don't need this trick for the covariance, but we do need to be careful. In fact we'll go back to the floor functions:

$$\begin{aligned}\mathrm{cov}(X_p,X_q) &= \mathbb{E}[X_pX_q] - \mathbb{E}[X_p]\mathbb{E}[X_q]\\ &= \tfrac{\lfloor n/(pq)\rfloor}{n} - \tfrac{\lfloor n/p\rfloor}{n}\tfrac{\lfloor n/q\rfloor}{n}\\ &\le \tfrac{1}{pq} - (\tfrac{1}{p}-\tfrac{1}{n})(\tfrac{1}{q}-\tfrac{1}{n})\\ &\le \tfrac{1}{n}(\tfrac{1}{p}+\tfrac{1}{q}).\end{aligned} \tag{3.5}$$

Part 2: It's now time to assess our strategy. Recall that X has the uniform distribution on $\{1,\dots,n\}$. The theorem claims precisely that

$$\mathbb{P}\left[|\nu(X) - \log\log n| \ge w(n)\sqrt{\log\log n}\right] \to 0 \tag{3.6}$$

as $n \to \infty$. Note that combining (3.1) with Lemma 3.2 will show that $\mathbb{E}[\nu(X)] \approx \log\log n$. We could then try to use Lemma 2.1 to find an upper bound for the left hand side of (3.6). This would mean calculating $\mathrm{var}(\nu(X))$, which is equal to

$$\mathrm{var}\left(\sum_{p\le n} X_p\right) = \sum_{p\le n} \mathrm{var}(X_p) + \sum_{\substack{p,q\le n\\ p\neq q}} \mathrm{cov}(X_p, X_q). \tag{3.7}$$

Unfortunately, the upper bound we get from (3.5) isn't strong enough for this, so we need to think of something better. Instead we will try to implement this strategy with with $M = n^{1/10}$ in place of n, which will mean there are less terms in (3.7) and hence better control. To this end set $\nu_M(X) = \sum_{p\le M} X_p$. We can get away with this strategy because $x \le n$ implies that x cannot have more than 10 prime factors that are greater than or equal to M, hence in fact

$$0 \le \nu(X) - \nu_M(X) \le 10. \tag{3.8}$$

In particular, it suffices to prove (3.7) with ν_M in place of ν. Let's aim for that.

By Lemma 3.2 and (3.1) we have

$$\begin{aligned}\mathbb{E}[\nu_M(X)] &= \sum_{p\le M} \left(\tfrac{1}{p} + \mathcal{O}(\tfrac{1}{n})\right)\\ &= \log\log M + M \times \mathcal{O}(\tfrac{1}{n})\\ &= \log\log n - \log 10 + \mathcal{O}(\tfrac{1}{n^{9/10}})\\ &= \log\log n + \mathcal{O}(1). \end{aligned} \tag{3.9}$$

In view of this, in fact we are going to end up putting $\nu_M(X) + \mathcal{O}(1)$ in place of $v(X)$, in (3.6), but that's also fine. Combining (3.9) with Theorem 2.2 we have that

$$\mathbb{P}\left[|\nu_M(X) - \log\log n + \mathcal{O}(1)| \ge w(n)\sqrt{\log\log n}\right] \le \frac{\mathrm{var}(\nu_M(X))}{w(n)^2 \log\log n}. \tag{3.10}$$

Using (3.7), but now only summing up to M, and putting in our estimates (3.4) and (3.5) we have

$$\begin{aligned}\mathrm{var}(\nu_M(X)) &\le \sum_{p\le M} \left(\tfrac{1}{p} + \mathcal{O}(\tfrac{1}{n})\right) + \sum_{\substack{p,q\le M\\ p\neq q}} \tfrac{1}{n}\left(\tfrac{1}{p} + \tfrac{1}{q}\right)\\ &= \log\log n + \mathcal{O}(1) + \frac{4M}{n}\sum_{p\le M} \tfrac{1}{p}\\ &\le \log\log n + \mathcal{O}(1) + \mathcal{O}(n^{-9/10})\left(\log\log n + \mathcal{O}(1)\right)\\ &= \log\log n + \mathcal{O}(1). \end{aligned} \tag{3.11}$$

Note that to deduce the second line of the above, for the first term we used the same calculation as in (3.9). For the second term we write the summation of p and q as $2\sum_{p\le M}\sum_{p<q}$, then note that the inner summation has at most M terms. To deduce the third line we use the same calculation as in (3.9), again.

Putting (3.11) into (3.10), we have that the right hand side of (3.10) tends to zero as $n \to \infty$ (because $w(n) \to \infty$). In view of (3.8) this implies that (3.6) also tends to zero, which completes the proof. ■

Equation (3.3) suggests that we could push this idea even further. If the events $\{p \text{ divides } X\}$ were actually independent (instead of just approximately) then we could apply the central limit theorem and deduce a connection between prime factorization and the normal distribution. This idea is the basis of a famous paper by Erdős and Kac (1940). You can also find it, alongside several other probabilistic proofs about prime numbers, within a short book by Kac (1959) that was recently reprinted.

Theorem 3.1 holds as $n \to \infty$. How well does that work in practice? From (3.8) and (3.10) we can see that our proof is allowing ν to vary by (at least) up to 10, and undoing the $\log\log$ gives $e^{e^{10}} \approx 9.4 \times 10^{9565}$. There are estimated to be about 10^{80} atoms within the universe, so its going to take a while to kick in. All is not lost though, because it takes a lot less that 10^{80} atoms to *think* about very large prime numbers and we might only be interested in a few of them anyway.

We might ask why the particular choice of $M = n^{1/10}$, within the proof of Theorem 3.1. Taking a larger value say $M = M_n = n^{1/2}$ would weaken the bound in (3.9) (which is bad) but would mean ν and ν_M were closer (which is good). Taking a smaller M_n, but still with $M_n \to \infty$, would have the reverse effects. There is a trade-off with no clear way to establish the best choice.

Lastly, what of the version of $\nu(n)$ where we count prime divisors with multiplicity, so that $\nu(3^2) = 2$ and $\nu(3) = 1$? This is harder to study but it behaves much like ν, at least in so far as properties like Theorem 3.1 are concerned. For that one, as far as I know, probability doesn't help and you'll need to consult a textbook on advanced number theory.

Exercise 3.3 Investigate the Erdős-Kac Theorem (using e.g. Wikipedia) and make your best effort at numerically testing its conclusion. Based on what you discover: if you didn't know it was true, do you think you have enough numerical evidence to believe it?

Chapter 4

Stirling's Formula

I'm sure you've seen Sterling's formula before. When you need to take a limit of some formula involving factorials, its usually the right tool for the job and there isn't much else available.

Theorem 4.1 (Stirling's Formula) *As $n \to \infty$ it holds that*

$$\frac{e^{-n}n^{n+1/2}}{n!} \to \frac{1}{\sqrt{2\pi}}. \tag{4.1}$$

Stirling's formula can be proved via calculus and some facts about convergence to the exponential function, for example based on the integrals $\int_0^\infty x^n e^{-x} = n!$ and the change of variables $t = \frac{x-n}{\sqrt{n}}$, but it needs quite a bit of care. One of the standard tools from probability theory can save a lot of work here. In any case, the appearance of $\frac{1}{\sqrt{2\pi}}$ should be enough to make you wonder if it's possible to prove Theorem 4.1 with probability.

Theorem 4.1 is concerned with convergence of real numbers. When we discuss convergence of real numbers $a_n \to a$, what we mean is that the value of a_n becomes close to the value of a_n. For random variables the situation is not so straightforward because random variables take different values with varied probabilities. In fact, there are many different ways in which we can make sense of convergence of random variables $X_n \to X$. They are normally introduced within third year probability courses. We only need one of them here.

Let X_n and X be real valued random variables. We say that the sequence (X_n) *converges in distribution* to X if

$$\mathbb{P}[X_n \le x] \to \mathbb{P}[X \le x] \text{ for all } x \in \mathbb{R} \text{ at which } \mathbb{P}[X = x] = 0.$$

We write this as $X_n \overset{\mathrm{d}}{\to} X$. It might seem unnatural to pick only on x such that $\mathbb{P}[X = x] = 0$ but it is necessary to do so: consider the example $X_n = \frac{1}{n}$, for which we obviously want $X_n \overset{\mathrm{d}}{\to} X = 0$, but $\mathbb{P}[X_n \le 0] = 0$ and $\mathbb{P}[X \le 0] = 1$. You've probably heard of the following theorem before.

Theorem 4.2 (Central Limit Theorem) *Suppose that (X_i) are independent, identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then as $n \to \infty$*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \overset{\mathrm{d}}{\to} N(0,1),$$

where $N(0,1)$ represents the normal distribution with mean zero and variance one.

Where to find it: You'll find a proof of the Central Limit Theorem somewhere within most maths degrees, sometimes early on without much rigour, sometimes late on with a full proof. ■

It is remarkable that the distribution of the X_i does not matter, and that in all cases (providing $\sigma^2 < \infty$) we obtain the normal distribution as the limit. However, for our purposes, we will only be interested in a particular example of the X_i. We need another standard result from probability too, one which pairs nicely with Theorem 4.2.

Lemma 4.3 (Moment Convergence) *Suppose that X_n, X are random variables and that $X_n \xrightarrow{\text{d}} X$. If we have $\sup_n \mathbb{E}[|X_n|^\alpha] < \infty$ where $\alpha \geq 1$ then*

$$\mathbb{E}[X_n^\beta] \to \mathbb{E}[X^\beta] \qquad \textit{and} \qquad \mathbb{E}[|X_n|^\beta] \to \mathbb{E}[|X|^\beta]$$

for all $\beta \in [1, \alpha)$.

Where to find it: Lemma 4.3 is really just an example of a more general result known as the Dominated Convergence Theorem; you'll find this in any course on Lebesgue integration, typically within second or third year. You could also look in e.g. Section 7.10 of the textbook by Grimmett and Stirzaker (2001). ■

Proof of Theorem 4.1: Let (X_i) be a sequence of independent, identically distributed Poisson(1) random variables, and let $S_n = \sum_{i=1}^n X_i$. We have $\mathbb{E}[X_i] = 1$ and $\text{var}(X_i) = 1$, so Theorem 4.2 gives that

$$\frac{S_n - n}{\sqrt{n}} \xrightarrow{\text{d}} N(0,1). \tag{4.2}$$

Let us write $Z_n = \frac{S_n - n}{\sqrt{n}}$. The plan is to apply Lemma 4.3 to $\mathbb{E}[|Z_n|]$ and use Theorem 4.2 to justify the application.

In the proof of Theorem 2.5 we noted that, for independent Poisson random variables, Poisson(a) + Poisson(b) has the same distribution as Poisson($a + b$). Hence $S_n \sim$ Poisson(n), which gives $\mathbb{E}[S_n] = n$ and $\text{var}(S_n) = n$. From this we have

$$\begin{aligned} \mathbb{E}[Z_n] &= \tfrac{1}{\sqrt{n}}(\mathbb{E}[S_n] - n) = 0 \\ \text{var}(Z_n) &= \frac{1}{n}\text{var}(S_n) = 1, \end{aligned}$$

and it follows immediately that $\sup_n \mathbb{E}[Z_n^2] = 1$. We may therefore apply Lemma 4.3 to equation (4.2), with $\beta = 1$ and $\alpha = 2$, and conclude that

$$\mathbb{E}[|Z_n|] \to \mathbb{E}[|Z|] \tag{4.3}$$

where Z has the $N(0,1)$ distribution. We will now calculate the two sides of (4.3). It will turn out that (4.3) gives us (4.1), which will complete the proof. Using that $Z \sim N(0,1)$ we have

$$\begin{aligned} \mathbb{E}[|Z|] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} |z| e^{-z^2/2}\, dz \\ &= 2\int_0^\infty \frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\, dz \\ &= 2\frac{1}{\sqrt{2\pi}} \left[-e^{-z^2/2} \right]_0^\infty \end{aligned}$$

$$= \sqrt{\frac{2}{\pi}}.$$

Note that for any $x \in \mathbb{R}$ we have $|x| = x - 2\mathbb{1}_{\{x<0\}}$. We will apply this to help us calculate $\mathbb{E}[|Z_n|]$, starting with $\mathbb{E}[|S_n - n|]$. We have

$$\begin{aligned}
\mathbb{E}[|S_n - n|] &= \mathbb{E}[S_n - n] - \mathbb{E}\left[(S_n - n)\mathbb{1}_{\{S_n - n < 0\}}\right] \\
&= 0 - \mathbb{E}\left[(S_n - n)\mathbb{1}_{\{S_n < n\}}\right] \qquad (4.4) \\
&= -2\sum_{s=0}^{n-1}(s-n)\frac{e^{-n}n^s}{s!} \\
&= -2e^{-n}\left(0 - \frac{n^1}{0!} + \sum_{s=1}^{n-1}\left(\frac{n^s}{(s-1)!} - \frac{n^{s+1}}{s!}\right)\right) \\
&= \frac{2e^{-n}n^n}{(n-1)!}. \qquad (4.5)
\end{aligned}$$

The third line uses that $S_n \sim \text{Poisson}(n)$. To deduce the fourth and fifth lines, we note that we have a telescoping sum in which the initial term is zero and only the final term from $s = n-1$ does not cancel. Multiplying the top and bottom of (4.5) by n and then dividing by $\sqrt{n}$, we obtain

$$\mathbb{E}[|Z_n|] = \frac{2e^{-n}n^{n+1/2}}{n!}.$$

From (4.3) we thus obtain

$$\frac{2e^{-n}n^{n+1/2}}{n!} \to \sqrt{\frac{2}{\pi}}.$$

Dividing both sides by 2 completes the proof. ■

We stated Theorem 4.1 for $n \in \mathbb{N}$, but in fact a more general result is true, where we allow $n \in (0, \infty)$ and replace $n!$ with $\Gamma(n+1)$, where $\Gamma(t)$ is the Gamma function,

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx.$$

If you try to prove Theorem 4.1 using calculus, even for just $n \in \mathbb{N}$, then you end up having to learn a lot about the Gamma function.

It's possible to prove the more general result using probability too, via essentially the same method as above but using the Gamma distribution in place of the Poisson distribution. You can find the details in Blyth and Pathak (1986).

Exercise 4.4 Use the central limit theorem and the Poisson distribution to show that

$$\lim_{n\to\infty} e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} = \frac{1}{2}.$$

Chapter 5

Partial Differential Equations

We'll need Brownian motion for Chapters 5 and 6. Brownian motion is usually introduced in a third year course on probability, so we need to spend a little time familiarising ourselves.

The discovery of Brownian motion has a distinguished place in the history of both science and mathematics. As with most great discoveries in the scientific world, many people discovered parts of what later became known as Brownian motion, using varying degrees of mathematical and experimental precision, at around the same time.

Brownian motion is named after the botanist Robert Brown who, in 1827, through a microscope, saw erratic movements being made by tiny pollen organelles floating on water. The cause of these movements was later explained in 1905 by Albert Einstein and the physicist Jean Perrin: the movements were caused by (the cumulative effect of) many individual water molecules hitting the tiny organelles. This realization provided the 'modern science' of the time with a key piece of evidence for the existence of atoms[1]. Around the same time, the american mathematician Norbert Wiener, building on earlier work of Louis Bachelier, developed a mathematical model for stock prices and independently discovered Brownian motion.

Today, Brownian motion is at the heart of many important models of the physical world. We will see some examples of this later on; for now our task is simply to construct the process.

Brown, Einstein and Perrin studied pollen movements on the surface of still water, meaning they observed movements in two (spatial) dimensions $\mathbb{R}^2$. Bachelier, by contrast, saw stocks prices moving up and down - in one dimension $\mathbb{R}$. In both cases the underlying principle is one of 'completely random' movement. We will restrict to the one dimensional case in this chapter.

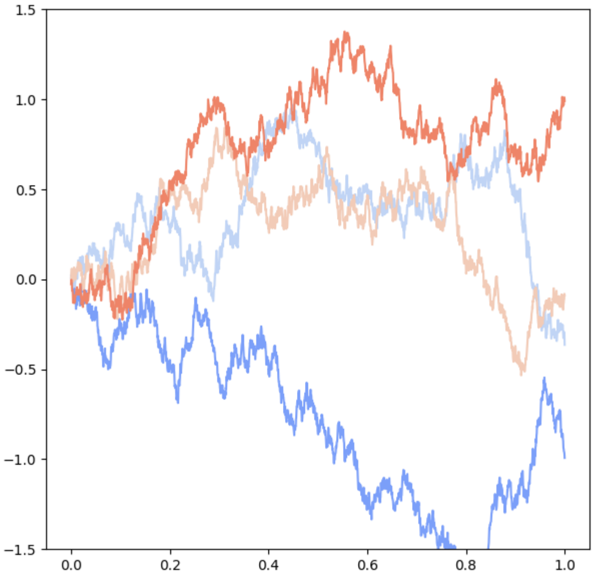
Theorem 5.1 (Existence of Brownian Motion) *Let $x \in \mathbb{R}$. There is a random continuous function $t \mapsto B_t$, defined for all $t \in [0,\infty)$ such that:*

1. *$B_0 = x$*
2. *For any $0 \le u \le t$, the random variable $B_t - B_u$ is independent of $\sigma(B_v\,;\, v \le u)$.*
3. *For any $0 \le u \le t$, the random variable $B_t - B_u$ has distribution $N(0, t-u)$.*

Further, any random continuous function which satisfies these conditions has the same distribution.

[1] At the time, scientists were not confident of the existence of atoms; there were other competing theories that had not yet been disproved.

Theorem 5.1 gives us some properties of (B_t) to work with, but from a practical point of view it is helpful to see some samples of the random continuous function $t \mapsto B_t$. Brownian motion can be set off from any point $x \in \mathbb{R}$, but we'll take four samples all starting from $B_0 = 0$.



The first thing you'll notice is that they are rough, not smooth; it is as though Brownian motion is always trying to turn corners in all directions. In fact, Brownian motion is not differentiable.

5.1 Heat flow

Consider a long thin metal rod. If, initially, some parts of the rod are hot and some are cold, then as time passes heat will diffuse through the rod: the differences in temperature slowly average out. Let the temperature of the metal in the rod at position x at time t be given by $u(t,x)$, where $t \in [0,\infty)$ represents time and $x \in \mathbb{R}$ represents space. Suppose that, at time 0, the temperature at the point x is $f(x) = u(0,x)$. It is well known that the *heat equation*

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} \tag{5.1}$$

with the initial condition

$$u(0,x) = f(x) \tag{5.2}$$

describes how the temperature $u(t,x)$ changes with time, from an initial temperature of $f(x)$ at site x. This equation has a close connection to Brownian motion, which we now explore.

Remark 5.2 In the world of PDEs the factor $\frac{1}{2}$ in (5.1) is not normally included, but in probability we tend to include it. The difference is just that time runs twice as fast (i.e. we substituted $2t$ in place of t), which isn't very important. In Chapter 6 we'll think about more sophisticated sorts of *time change*.

If we start Brownian motion from $x \in \mathbb{R}$, then $B_t = B_0 + (B_t - B_t) \sim x + N(0, t-0) \sim N(x,t)$, so we can write down the probability density function of B_t,

$$\phi_{t,x}(y) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right).$$

Lemma 5.3 *$\phi_{t,x}(y)$ satisfies the heat equation* (5.1)*.*

PROOF: Note that if any function $u(t,x)$ satisfies the heat equation, so does $u(t,x-y)$ for any value of y. So, we can assume $y=0$ and need to show that

$$\phi_{t,x}(0)=\frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{x^2}{2t}\right)$$

satisfies (5.1). This is an exercise in partial differentiation. Using the chain and product rules:

$$\begin{aligned}\frac{\partial\phi}{\partial t}&=\frac{\frac{-1}{2}}{\sqrt{2\pi t^3}}\exp\left(-\frac{x^2}{2t}\right)+\frac{1}{\sqrt{2\pi t}}\frac{-x^2(-1)}{2t^2}\exp\left(-\frac{x^2}{2t}\right)\\&=\frac{1}{\sqrt{2\pi}}\left(\frac{x^2}{2t^{5/2}}-\frac{1}{2t^{3/2}}\right)\exp\left(-\frac{x^2}{2t}\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\phi}{\partial x}&=\frac{1}{\sqrt{2\pi t}}\frac{-2x}{2t}\exp\left(-\frac{x^2}{2t}\right)\\&=\frac{-x}{\sqrt{2\pi t^3}}\exp\left(-\frac{x^2}{2t}\right)\end{aligned}$$

so that

$$\begin{aligned}\frac{\partial^2\phi}{\partial x^2}&=\frac{-1}{\sqrt{2\pi t^3}}\exp\left(-\frac{x^2}{2t}\right)+\frac{-x}{\sqrt{2\pi t^3}}\frac{-2x}{2t}\exp\left(-\frac{x^2}{2t}\right)\\&=\frac{1}{\sqrt{2\pi}}\left(\frac{x^2}{t^{5/2}}-\frac{1}{t^{3/2}}\right)\exp\left(-\frac{x^2}{2t}\right).\end{aligned}$$

Hence, $\frac{\partial\phi}{\partial t}=\frac{1}{2}\frac{\partial^2\phi}{\partial x^2}$. ■

We can use Lemma 5.3 to give a physical explanation of the connection between Brownian motion and heat diffusion. We define

$$w(t,x)=\mathbb{E}_x[f(B_t)] \tag{5.3}$$

That is, to get $w(t,x)$, we start a particle at location x, let it perform Brownian motion for time t, and then take the expected value of $f(B_t)$.

Theorem 5.4 *$w(t,x)$ satisfies the heat equation* (5.1) *and the initial condition* (5.2)*.*

Before we give the proof, let us discuss the physical interpretation of this result. Within our metal rod, the metal atoms have fixed positions. Atoms that are next to each other transfer heat between each other, in random directions. If we could pick on an individual 'piece' of heat and watch it move, it would move like a Brownian motion. Since there are lots of little pieces of heat moving around, and they are *very* small, when we measure temperature we only see the average effect of all the little pieces, corresponding to $\mathbb{E}[\ldots]$.

We should think of the Brownian motion in (5.3) as running in reverse time, so as it tracks (backwards in time) the path through space that a typical piece of heat has followed. Then, after running for time t, it looks at the initial condition to find out how much heat there was initially that its eventual location.

Proof: We have $B_0 = x$, so $w(0,x) = \mathbb{E}_x[f(B_0)] = \mathbb{E}_x[f(x)] = f(x)$. Hence the initial condition (5.2) is satisfied. We still need to check (5.1). To do so we will allow ourselves to swap $\int$s and partial derivatives – strictly, this requires an application of the dominated convergence theorem, which you can find within any course on Lebesgue integration. We have

$$w(t,x) = \mathbb{E}_x[f(B_t)] = \int_{-\infty}^{\infty} f(y)\phi_{t,x}(y)\,dy$$

so, by Lemma 5.3,

$$\begin{aligned}\frac{\partial w}{\partial t} &= \frac{\partial}{\partial t}\int_{-\infty}^{\infty} f(y)\phi_{t,x}(y)\,dy\\ &= \int_{-\infty}^{\infty} f(y)\frac{\partial}{\partial t}\phi_{t,x}(y)\,dy\\ &= \int_{-\infty}^{\infty} f(y)\frac{1}{2}\frac{\partial^2}{\partial x^2}\phi_{t,x}(y)\,dy\\ &= \frac{1}{2}\frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty} f(y)\phi_{t,x}(y)\,dy\\ &= \frac{1}{2}\frac{\partial^2 w}{\partial x^2}\end{aligned}$$

as required. ∎

Leading on directly from this connection, it becomes possible to relate properties of heat diffusion to properties of Brownian motion. In this area there are major results where the only known proofs are via probability. One example is the 'hot spots' conjecture, which loosely states that once heat diffusion has applied to a finite object for a long period of time, the hottest and coldest spots will be near the boundary of the object. This is natural: nearby hot and cold regions annihilate one another, and there is more scope for that to happen in the middle than at the edges. The book of Burdzy (2014) contains more details.

5.2 Travelling waves

You are probably expecting me to discuss the wave equation, but I'm not going to do that. The wave equation describes things like water waves, sound waves and electromagnetic waves i.e. waves that transport energy and are collectively known as 'mechanical waves'. We are interested in a different type of wave, which tends to describe the spread of some invasive quantity through space. Two examples illustrate the idea: selectively advantageous genes displace weaker genes and consequently spread out as organisms simultaneously move and reproduce; in combustion, a wavefront emerges at the leading edge of a fire that burns through a region of fuel. The equation we are interested in here is

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + \beta u(u-1), \tag{5.4}$$

known as the Fischer-Kolmogorov-Petrovsky-Piskunov equation, or FKPP for short. Here, $u(t,x)$ is a function of time $t \in [0,\infty)$ and space $x \in \mathbb{R}$, and $\beta > 0$ is a constant. Compared to the heat equation (5.1), there's an extra term on the right of (5.4). This changes the behaviour drastically: it creates a force pushing upwards from 0 towards 1. The strength of this upwards force depends on the current value of $u(t,x)$. The push upwards is non-zero for $u\in(0,1)$, strongest at $u=\frac{1}{2}$ and vanishing at $u=0,1$, and if this force is strong enough then it will lead to $u\approx 0$ rising to $u\approx 1$ over time. We can recover (5.1) by setting $\beta=0$.

We will focus on the initial condition $f(x)=\mathbb{1}_{x\geq 0}$, which is often known as the Heaviside function. It is possible to extend our analysis below to cover general initial conditions, but the Heaviside function will be enough to illustrate the key idea.

We can't re-use our method of proof from Section 5.1 because there is no easy explicit formula to take the place of Lemma 5.3. Instead we need a more sophisticated argument, as well as a a more sophisticated type of particle motion. For this we'll need the *memorylessness* of the exponential distribution, which you sometimes see an example (of conditional probability) in second year, but which becomes central in much of continuous time probability theory.

Lemma 5.5 *Let X have the Exponential(λ) distribution, where $\lambda>0$. Then $\mathbb{P}[X\geq t+s\,|\,X\geq t]=\mathbb{P}[X\geq s]$, for all $s,t\geq 0$.*

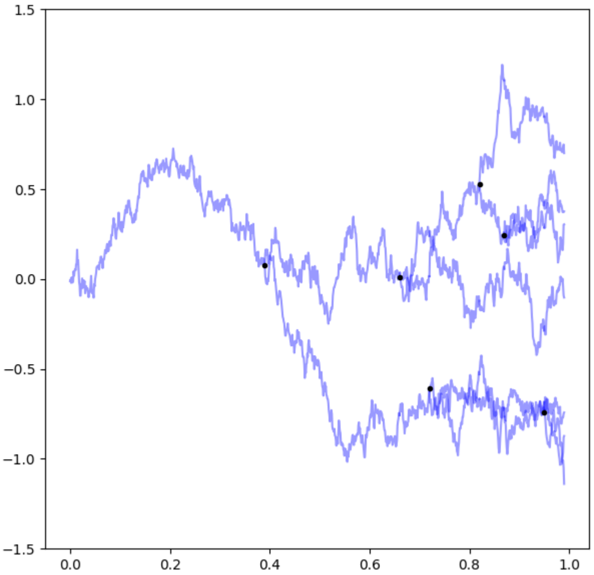
PROOF: Recall that the p.d.f. of the exponential distribution is $f(t)=\lambda e^{-\lambda t}$. Thus $\mathbb{P}[X\geq s]=\int_s^\infty \lambda e^{-\lambda u}\,du=e^{-\lambda s}$ and

$$\mathbb{P}[X\geq t+s\,|\,X\geq t]=\frac{\mathbb{P}[X\geq t+s,X\geq t]}{\mathbb{P}[X\geq t]}=\frac{\mathbb{P}[X\geq t+s]}{\mathbb{P}[X\geq t]}=\frac{e^{-\lambda(t+s)}}{e^{-\lambda t}}=e^{-\lambda s}$$

as required. ■

The point is that, if X is a random time with Exponential(λ) distribution, and we are waiting for X to happen, but we have already waited time $[0,t]$ during which X has not happened, then the remaining time to wait is still an Exponential(λ). You'll often hear of waiting for buses as an example of this, but it is a bad example: the exponential distribution is (in terms of entropy) the most random way in which an event might occur in time, whereas buses tend to arrive when the timetable says that a bus is due, and that is about as far from random as you can get. Nonetheless Lemma 5.5 is very useful for modelling random events.

We also need to introduce *branching Brownian motion* $\mathbf{B}=(B_t^{(N_t)})$. Here's a picture, with description below:



The process $\mathbf{B}$ starts at time 0 with a single particle at the origin. It moves according to Brownian motion. The particle is equipped with an exponential random variable, which is usually called an *exponential clock* in this context. When that exponential time runs out our particle 'dies' and is replaced by two new particles. This is called a *branching event.* The two new particles both start in the same location as where the old one died. Each new particle moves independently as a Brownian motion and carries its own (independent) exponential clock; when its time is up it dies and is replaced by two more, and so on. The quantity N_t counts the number of particles alive in $\mathbf{B}$ at time t.

Using the memoryless property of the exponential distribution from Lemma 5.5, plus some additional work, its possible to show that the whole process $\mathbf{B}$ is memoryless in a similar way. More precisely, if we want to describe the future behaviour of $\mathbf{B}$, we only need to be told the current state, and then the rules above determine the (random!) future. This idea is called the *Markov property.* It is one of the most powerful tools in probability and you'll certainly come across it at some point within probability courses, if you haven't already. It may sound intuitive but it takes quite a bit of hard work to make this idea mathematically rigorous.

Theorem 5.6 (McKean's solution to FKPP) *Let $\boldsymbol{B} = (B_t^{(n)})_{n=1}^{N_t}$ be a branching Brownian motion with branching rate $\beta \geq 0$, started at the origin. Let*

$$M_t = \max_{n=1,\dots,N_t} B_t^{(n)}$$

denote the maximal position of particles of $\boldsymbol{B}$, at time t. Then

$$u(x,t) = \mathbb{P}[M_t \leq x]$$

solves (5.4)*, subject to the initial condition $u(x,0) = \mathbb{1}_{\{x \geq 0\}}$.*

Sketch of Proof: We'll sketch a proof but we won't be able to fully justify some of the approximations. Our strategy in this case is to calculate $\frac{\partial u}{\partial t}$ directly, where $u(t,x) = \mathbb{P}[M_t \leq x]$.

Let's begin by thinking about what branching Brownian motion might do in the time interval $[0,\delta]$. We start at time $t = 0$ with a single particle at the origin. The probability of seeing a branching event during time $[0,\delta]$ is $\int_0^\delta \beta e^{-\beta t}\,dt = 1 - e^{-\beta\delta} \approx \beta\delta$, where we have used the approximation $1 - e^{-x} \approx x$ when $x \approx 0$. A similar calculation shows that the chance of seeing two or more branching events during $[0,\delta]$ is $\mathcal{O}(\delta^2)$, which will turn out to be negligibly small, so lets ignore that. Let us write E_δ for the event that a single branching event occurs during $[0,\delta]$.

- If there is no branching event during $[0,\delta]$ then, conditionally on this event we have

  $$\mathbb{P}[M_{t+\delta} \leq x] = \mathbb{P}[M_t \leq x - B_\delta] = \mathbb{E}[u(t, x - B_\delta)].$$

  Here, (B_t) is a Brownian motion started from zero, representing the moving of the (non-branching) particle during $[0,\delta]$. Note that the justification for the last inequality is similar to that of (2.10) and uses that B_δ and M_t are independent – strictly, it is an application of Fubini's theorem.

- If a branching event occurs during $[0,\delta]$ then as soon as event occurs, we have two particles, each of which will now behave like an independent copy of branching Brownian motion.

Therefore, conditionally on this event we have

$$\mathbb{P}[M_{t+\delta} \leq x] \approx \mathbb{P}[B_\delta + M_t \leq x]\mathbb{P}[B'_\delta + M'_t \leq x] = \mathbb{P}[B_\delta + M_t \leq x]^2 = \mathbb{E}[u(t, x - B_\delta)]^2.$$

Here (M'_t) is an independent copy of (M_t), and (B_t) and (B'_t) are correlated Brownian motions, which are starting from zero and are equal up until the branching event happens, and independent thereafter.

Putting these two cases together, we have

$$u(x, t+\delta) = \mathbb{P}[M_{t+\delta} \leq x] = (1 - \beta\delta)\mathbb{E}[u(t, x - B_\delta)] + (\beta\delta)\mathbb{E}[u(t, x - B_\delta)]^2$$

and hence

$$u(x, t+\delta) - u(x, t) = (1 - \beta\delta)\mathbb{E}\left[u(t, x - B_\delta) - u(t, x)\right] + \beta\delta\left(\mathbb{E}[u(t, x - B_\delta)]^2 - u(t, x)\right).$$

Dividing through by δ we obtain

$$\frac{u(x, t+\delta) - u(x, t)}{\delta} = (1 - \beta\delta)\mathbb{E}\left[\frac{u(t, x - B_\delta) - u(t, x)}{\delta}\right] + \beta\left(\mathbb{E}[u(t, x - B_\delta)]^2 - u(t, x)\right). \quad (5.5)$$

We now send $\delta \to 0$. Strictly, we should check that u is differentiable (and hence also continuous) but we'll omit this part. The left hand side will converge to $\frac{\partial u}{\partial t}$. In the rightmost term, we will see $B_\delta \to 0$, because the Brownian motion is continuous and started as zero, so this term converges to $\beta(\mathbb{E}[u(t, x)]^2 - u(t, x)) = \beta(u^2 - u)$.

We will use Theorem 5.4 for the remaining term. Using $\mathbb{E}_x$ to denote Brownian motion started from x, we have

$$\mathbb{E}\left[\frac{u(t, x - B_\delta) - u(t, x)}{\delta}\right] = \mathbb{E}_x\left[\frac{g(B_\delta) - g(B_0)}{\delta}\right] \to \frac{\partial g}{\partial t} = \frac{\partial^2 g}{\partial^2 x} = \frac{\partial^2 u}{\partial x^2} \quad (5.6)$$

where $g(x) = u(t, x)$. Note that in the first step we have used that $B_\delta \sim N(0, \delta)$ has the same distribution as $-B_\delta$. Putting (5.6) together with our calculations above, after sending $\delta \to 0$ in (5.5) we obtain that $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \beta(u^2 - u)$, as required. ∎

I've worked on this area, so I'll tell you a bit about another example, one that I was involved in. The Allen-Cahn equation (with a particular choice of parameters) is the PDE

$$\frac{du}{dt} = \frac{\partial^2 u}{\partial x^2} - C(u^3 - u) \quad (5.7)$$

where $C > 0$. It isn't obvious from (5.7), but this equation can be used to describe a travelling wave, moving across a two dimensional surface, in which the speed of movement (of the wave) depends on how sharply curved the wavefront is. Note the change of sign compared to (5.4) which means that, if we tried to use our method of proof for Theorem 5.6 in this case, we would end up trying to define a branching Brownian motion with negative branching rate. That doesn't exist – at first sight the method of proof appears to fall apart here.

However, it is possible to make a connection between (5.7) and branching Brownian motion. I won't explain the details here, but they are written up in Etheridge et al. (2017). We were interested in (5.7) for reasons related to modelling natural selection, in a particular scenario where two populations of animals border each other and can interbreed, but when they mix they create 'hybrid' children that are less genetically fit than non-hybrids.

This situation exists, for example, within mainland Europe in the house mouse population, with a curved interface that runs right across Germany and down to the Black Sea. It is (coincidentally) similar to the path of the Iron Curtain[2], and for this reason the mouse version is often known as the Squeaky Curtain.

There are various differences between the two sub-populations of mice but they are easy to distinguish by eye: the eastern variant is smaller and browner; the western variant is larger and greyer. The hybrids occur only within a narrow interface between these two sub-populations. The hybrids are less fit, so the interface in which they exist does not expand. Making the connection to Brownian motion corresponds to relating the movement of the interface to the genealogical tree of the population.

Exercise 5.7 Suppose that our branching Brownian motion splits into $k \geq 2$ particles at each branching event, rather than just two. Which equation does this correspond to, in place of (5.4)? What about if k becomes random?

[2] The Iron Curtain was the path of national borders that separated Western Europe from the Soviet Union during the cold war.

Chapter 6

Complex analysis

In Chapter 5 we thought about Brownian motion in one dimension. In this chapter we will think about Brownian motion in two dimensions. Fortunately the definition doesn't need much modification. If $(B^1_t)_{t\geq 0}$ and $(B^2_t)_{t\geq 0}$ are a pair of independent one dimensional Brownian motions then we define

$$B_t = \big(B^1_t, B^2_t\big)$$

and then $(B_t)_{t\geq 0}$ is a two dimensional Brownian motion. The same idea works in dimensions three and higher, but we'll need to focus on $d=2$ here. We would prefer to work in $\mathbb{C}$ than $\mathbb{R}^2$, for which purposes we will write

$$B_t = B^1_t + iB^2_t,$$

where of course $i=\sqrt{-1}$. This is know as a *complex Brownian motion.*

Let $\tau : [0,\infty) \to [0,\infty)$ be continuous and strictly increasing (and perhaps random), with $\tau(0)=0$ and $\lim_{t\to\infty}\tau(t)=\infty$. We say that the process $t\mapsto B_{\tau(t)}$ is a *time changed* Brownian motion and the function τ is a time change. The point is that the function τ allows time to run faster or slower. For example $t\mapsto B_{2t}$ has time running twice as fast as normal, and this is $\tau(t)=2t$. Of course, we can use this idea on any stochastic process.

We need to introduce some structure surrounding complex functions. You might have seen some of it before, or perhaps all of it if you've already studied complex analysis. For $z\in\mathbb{C}$ and $\epsilon>0$ let $\mathcal{B}_\epsilon(z)=\{w\in\mathbb{C}\,;\,|z-w|\leq\epsilon\}$. We say that a subset U of $\mathbb{C}$ is *open* if for any $z\in C$ there exists $\epsilon>0$ such that $\mathcal{B}_\epsilon(z)\subseteq U$.

Convergence in $\mathbb{C}$ works exactly as you would expect: we say $z_n\to z$ if $\Re(z_n)\to\Re(z)$ and $\Im(z_n)\to\Im(z)$. We say that a function $f:U\to\mathbb{C}$ is *complex-differentiable* at $z_0\in U$, or just differentiable, if the limit

$$f'(z) = \lim_{z\to z_0}\frac{f(z)-f(z_0)}{z-z_0} \tag{6.1}$$

exists. We use both Newton and Leibniz notation: $f'=\frac{df}{dz}$. It's the natural analogue of real differentiation, but note that in $\mathbb{C}$ we can have z approaching z_0 in many different ways (for example, it might spiral in) and that (6.1) requires that the same limit is obtained regardless of the way in which $z\to z_0$. Consequently (6.1) is a stronger condition than in $\mathbb{C}$ than $\mathbb{R}$.

Provided that we deal with nice enough functions then complex differentiation behaves as you'd expect. For example $\frac{d}{dz}z^2=2z$ and $\frac{d}{dz}\sin z=\cos z$, even $\frac{d}{dz}\log z=\frac{1}{z}$, for $z\neq 0$. There are some things that don't work too, and we don't have anything like enough time to explore this area, so I'm afraid you'll have to take calculations about complex differentiation on trust.

From the probabilistic side, the main ingredient of this chapter is the following result. A function $f:\mathbb{C}\to\mathbb{C}$ that is defined and differentiable on the whole of $\mathbb{C}$ is said to be an *entire* function.

Theorem 6.1 (Lévy's Conformal Invariance) *Let (B_t) be a complex Brownian motion and let $f:\mathbb{C}\to\mathbb{C}$ be a non-constant entire function. Define $W_t=f(B_t)$. Then there exists a time change $\tau:[0,\infty)\to[0,\infty)$ such that $(W_{\tau(t)})$ is a complex Brownian motion.*

You might see Theorem 6.1 in a masters level probability course, but it probably won't appear before that. In that context it is not particular hard to prove, but it does take a lot of mathematics to reach a point where you can even approach it. You will often find a one-dimensional version of the result, called the Dubins-Schwarz Theorem, within a third or fourth year course on stochastic processes. We won't discuss the proof here.

The reason for including this chapter is that many important results in complex analysis can be proved with the help of Theorem 6.1. We'll see some example of this below, one of which is one of the most important results in mathematics, but we won't be able to do much more than scratch the surface of this area. We'll need the following result on our way.

Lemma 6.2 (Recurrence of Brownian motion) *Let (B_t) be a two dimensional Brownian motion. Then for any $z_0\in\mathbb{C}$ and $\epsilon>0$, it holds that $\mathbb{P}[\exists t\in(0,\infty)$ with $B_t\in\mathcal{B}_\epsilon(z_0)]=1$.*

Sketch of Proof: Note that this proof is only a sketch – it will have gaps, but it should illustrate a (very) plausible strategy. We first need a one dimensional version of the same result, from which Theorem 6.1 will provide the extension to $\mathbb{C}$.

We will show that if (U_t) is a one dimensional Brownian motion then for any $r>0$ we have

$$\mathbb{P}[\exists t\in(0,\infty)\text{ with }U_t>r]=1. \tag{6.2}$$

We claim that this is true regardless of the starting point $U_0=x_0\in\mathbb{R}$. Note that we have

$$\begin{aligned}\mathbb{P}[U_t>r]&=\int_r^\infty\frac{1}{\sqrt{2\pi t}}e^{-(x-x_0)^2/2t}\,dx\\&=\int_{(r-x_0)/\sqrt{t}}\frac{1}{\sqrt{2\pi t}}e^{-y^2t/2t}\sqrt{t}\,dy\\&=\int_{(r-x_0)/\sqrt{t}}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,dy\\&\to\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-y^2/2}\,dy=\mathbb{P}[N(0,1)\geq 0]=\frac{1}{2}\qquad\text{as }t\to\infty.\end{aligned}$$

Note that in the second line we make the substitution $y=(x-x_0)/\sqrt{t}$. Take $n\in\mathbb{N}$ and consider the events $E_n=\{U_n>r\}$. From above we have $\mathbb{P}[E_n]\approx\frac{1}{2}$ for large n. If the (E_n) were independent of each other then this would be enough to guarantee that at least one (in fact, infinitely many) of the E_n would occur, and this would prove (6.2). The (E_n) are not independent but, if n_2-n_1 is large then E_{n_2} and E_{n_1} are approximately independent of each other; this is because in the time in between n_1 and n_2 Brownian motion does lots of random movement and (to a great extent) forgets about its past. Instead, we should look at $\{E_{n_1},E_{n_2},\ldots\}$ where $n_1\ll n_2\ll\ldots$ and deduce (6.2) from that.

Our next task it to upgrade (6.2) into two dimensions. For this we will use Theorem 6.1. The first step is to construct a map that takes $\mathcal{B}_\epsilon(z_0)$ to $\mathbb{H} = \{z \in \mathbb{C}\,;\, \Im(z) > 0\}$. This comes in two steps.

- First, apply the map $z \mapsto \frac{z-z_0}{\epsilon}$, which takes $\mathcal{B}_\epsilon(z_0)$ to the unit disc $\{z \in \mathbb{C}\,;\, |z| < 1\}$. As you might expect, the derivative of this map is $\frac{1}{\epsilon}$, which is never zero.
- Second, apply the map $z \mapsto \frac{z-1}{z+1}$, which takes $\{z \in \mathbb{C}\,;\, |z| < 1\}$ to $\{z \in \mathbb{C}\,;\, \Re(z) > 0\}$. To see this, note that $\Re(z) > 0 \Leftrightarrow |z-1| > |z-(-1)| \Leftrightarrow \frac{|z-1|}{|z+1|} < 1$. The derivative of this map is $\frac{2}{(z+1)^{-2}}$, which is non-zero, except where $z = -1$, where the map is not defined.

Let ϕ be the composition of these two steps, so ϕ takes $\mathcal{B}_\epsilon(z_0)$ to $\mathbb{H}$. We've shown that the derivative, in each step, exists and is non-zero. Consequently ϕ is differentiable and non-constant.

The second stage of the map ϕ was undefined at the single point $z = -1$. For this reason ϕ doesn't quite satisfy the conditions we stated for Theorem 6.1. There's a way to get around this difficulty[1] and apply Theorem 6.1 anyway, so let us take (B_t) to a Brownian motion and apply Theorem 6.1 to (B_t) and ϕ. Then, with $W_t = \phi(B_t)$, there exists a time change τ such that (W_t) is a Brownian motion. By our choice of ϕ we have that $B_t \in \mathcal{B}_\epsilon(z_0)$ if and only if $W_t \in \mathbb{H}$, so if we can prove that (W_t) enters $\mathbb{H}$ then we are done.

We have that $(W_{\tau(t)})$ is a two dimensional Brownian motion, with initial point $W_0 = \phi(B_0)$, which might be anywhere in $\mathbb{C}$. In particular, the one real part of $(W_{\tau(t)})$ is a one dimensional Brownian motion, to which (6.2) applies, hence $W^1_{\tau(t)}$ is greater than zero at some point in time. At that time, $W_{\tau(t)} \in \mathbb{H}$, as required. ■

Given the method of proof above you might wonder if we can transfer Lemma 6.2 into three dimensions and higher. The answer is no. In dimensions three and above Brownian motion still has a positive probability of entering any ball $\mathcal{B}_\epsilon(z_0)$, but that probability becomes less than one. In three dimensions there is a lot more space to get lost inside than in two, so much so that three dimensional Brownian motion eventually disappears off to infinity and does not return. We won't be able to explore the properties of Brownian motion any further here, although this is a very interesting topic to study.

6.1 Entire functions

Our next result will be found within a first course on complex analysis, where it is normally obtained as a consequence of one of the major results of contour integration, which is a version of integration adapted to $\mathbb{C}$. It is also possible to obtain it as a consequence of Theorem 6.1, as follows.

Theorem 6.3 (Liouville's Theorem) *Let f be an entire function. If f is bounded then f is constant.*

PROOF: Suppose that f is a bounded entire function and that f is non-constant. Let (B_t) be a complex Brownian motion. By Theorem 6.1 there exists a time change τ such that $(W_{\tau(t)})$ is a complex Brownian motion, where $W_t = f(B_t)$. Since f is bounded we have $M \in \mathbb{R}$ such that

[1] They key idea is that, despite Lemma 6.2, two dimension Brownian motion (B_t) satisfies $\mathbb{P}[\exists t \in (0,\infty), B_t = z] = 0$ for all $z \in \mathbb{C}$. Consequently it is never equal to the single point at which ϕ is not defined.

$|f(z)| \leq M$ for all $z \in \mathbb{C}$. Hence $(W_{\tau(t)})$ remains within $\mathcal{B}_0(M)$ for all time. This contradicts Lemma 6.2. ■

It is possible to go a lot further than Theorem 6.3. Picard's Little Theorem says that a non-constant entire function f will have the whole of $\mathbb{C}$ in its range *except possibly a single value*. Both possibilities happen: the entire function $z \mapsto e^z$ (which unsurprisingly has e^z has its derivative) never takes the value 0, and the entire function $z \mapsto z$ (which has derivative 1) takes all values.

Picard's Little Theorem also has a proof based on complex Brownian motion. The argument has much in common with our proof of Theorem 6.3. It uses Theorem 6.1 plus a property about the behaviour of the paths of complex Brownian motion, but a much more advanced property than Lemma 6.2 is required. There is also a 'Picard's Great Theorem' and, more generally, there is a branch of complex analysis called Nevanlinna Theory that studies the values taken by entire functions. The main theorems in this area are deep results that describe how often each value within the range of an entire function is taken. They were first proved using complex analysis but also have probabilistic proofs.

6.2 The fundamental theorem of algebra

You already know this one:

Theorem 6.4 (Fundamental Theorem of Algebra) *Let $f(z) = \sum_{i=1}^n c_i z^i$ be a polynomial function with coefficients $c_i \in \mathbb{C}$. Then there exists $z \in \mathbb{C}$ such that $f(z) = 0$.*

There is a second part to this result, which says that f has precisely n roots after accounting for multiplicity. Given the above, that second part is straightforward to prove by induction, so we will omit it.

The fundamental theorem of algebra has a long history. During the 17th century mathematicians began to realize that something approximating Theorem 6.4 would be true. The road towards guessing the right result and then proving it was a long one: Leibnitz published an (incorrect) counterexample in 1702 and over the next century incomplete proofs were given by D'Alembert, Euler, Lagrange, Laplace and others. In those days analysis (and its axiomatic foundations) had not yet been made rigorous, meaning that 'proofs' were more prone to later having gaps found in than today. By today's standards the first known proof was given by Argand in 1806, and two more proofs were given by Gauss in 1816. You can take some comfort in this sort of thing when you discover gaps in your own proofs – the worlds best mathematicians did it that way for centuries. The sort of rigour that we enforce today has only been around for about 150 years.

Proof of Theorem 6.4: The function $f : \mathbb{C} \to \mathbb{C}$ is entire, and the derivative is what you think it ought to be. Let (B_t) be a complex Brownian motion and set $W_t = f(B_t)$. By Theorem 6.1 there exists a time change τ such that $(W_{\tau(t)})$ is a complex Brownian motion. By Lemma 6.2, setting $\epsilon = \frac{1}{n}$, for all $n \in \mathbb{N}$ there exists $s_n = \tau(t_n) \in \mathbb{R}$ such that

$$W_{s_n} = f(B_{s_n}) \in \mathcal{B}_{1/n}(0). \tag{6.3}$$

It is straightforward to show that $|f(z)| \to \infty$ as $|z| \to \infty$. If the sequence (B_{s_n}) was unbounded then we would have $f(B_{s_n}) \to \infty$, which would contradict (6.3), so the sequence (B_{s_n}) must be bounded. By the Heine-Borel theorem (applied first to real, then imaginary parts) it has a convergent subsequence, which we will denote by $B_{\hat{s}_n} \to z$. By (6.3) we have $f(B_{\hat{s}_n}) \to 0$ and therefore $f(z) = 0$. ■

Bibliography

N. Alon and J. H. Spencer. *The Probabilistic Method.* Wiley-Interscience, 2008.

C. R. Blyth and P. K. Pathak. A note on easy proofs of Stirling's theorem. *The American Mathematical Monthly*, 93(5):376, 1986.

K. Burdzy. *Brownian Motion and its Applications to Mathematical Analysis.* Springer International Publishing, 2014.

P. Erdős and M. Kac. The Gaussian law of errors in the theory of additive number-theoretic functions. *American Journal of Mathematics*, 62:738–742, 1940.

P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. *Infinite and Finite Sets (to Paul Erdős on his 60th birthday)*, 1975.

A. Etheridge, N. Freeman, and S. Penington. Branching brownian motion, mean curvature flow and the motion of hybrid zones. *Electronic Journal of Probability*, 22(none), 2017.

J. Fabius. A probabilistic example of a nowhere analytic c^∞-function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(2):173–174, 1966.

J. A. Goldstein. Some applications of the law of large numbers. *Boletim da Sociedade Brasileira de Matemática*, 6(1):25–38, 1975.

G. Grimmett and D. Stirzaker. *Probability and random processes.* Oxford University Press, 2001.

G. H. Hardy and S. Ramanujan. The normal number of prime factors of a number n. *The Quarterly Journal of Pure and Applied Mathematics*, 48:76–92, 1917.

M. A. Henning and A. Yeo. 2-colorings in k-regular k-uniform hypergraphs. *European Journal of Combinatorics*, 34(7):1192–1202, 2013.

B. Jessen and A. Wintner. Distribution functions and the riemann zeta function. *Transactions of the American Mathematical Society*, 38(1):48–88, 1935.

M. Kac. *Statistical Independence in Probability, Analysis and Number Theory.* Dover Publications, Incorporated, reprinted in 2018, 1959.

H. McKean. *Probability: The Classical Limit Theorems.* Cambridge University Press, 2014.

P. Turán. On a theorem of Hardy and Ramanujan. *Journal of the London Mathematical Society*, 9:274–276, 1934.